Rejection sampling; introduction to the normal model

Dr. Olanrewaju Michael Akande

Jan 29, 2020

Outline

- Sampling methods
  - Rejection sampling
  - Importance sampling
- The univariate normal model
  - Motivating example
  - The normal distribution
    - Properties
    - Likelihood
  - Inference for mean, conditional on variance
  - Conjugacy
  - Noninformative and improper priors

Sampling methods

Rejection sampling

- Setup:
  - $p(\theta)$ is some density we are interested in sampling from;
  - $p(\theta)$ is tough to sample from but we are able to evaluate $p(\theta)$ as a function at any point; and
  - $g(\theta)$ is some proposal distribution or importance sampling distribution that is easier to sample from.
- Two key requirements:
  - $g(\theta)$ is easy to sample from; and
  - $g(\theta)$ is easy to evaluate at any point as is the case for $p(\theta)$.
- Usually, the context is one in which $g(\theta)$ has been derived as an analytic approximation to $p(\theta)$; and the closer the approximation, the more accurate the resulting Monte Carlo analysis will be.

Rejection sampling

- Procedure:

  1. Define $w(\theta) = p(\theta)/g(\theta)$.
  2. Assume that $w(\theta) = p(\theta)/g(\theta) < M$ for some constant M. If $g(\theta)$ represents a good approximation to $p(\theta)$, then M should not be too far from 1.
  3. Generate a *candidate* value $\theta \sim g(\theta)$ and **accept** with probability $w(\theta)/M$: if accepted, θ is a draw from $p(\theta)$; otherwise **reject** and try again.
     Equivalently, generate $u \sim U(0,1)$ independently of θ. Then **accept** θ as a draw from $p(\theta)$ if, and only if, $u < w(\theta)/M$.

- For those interested, the proof that all accepted θ values are indeed from $p(\theta)$ is on the next slide. We will not spend time on it.

- Clearly, we need M for this to work. However, in the case of truncated densities, we actually have M.

Proof for simple accept/reject

- We need to show that all accepted θ values are indeed from $p(\theta)$. Equivalently, show that $f(\theta|u < w(\theta)/M) = p(\theta)$.
- By Bayes' theorem,

$$f(\theta|u < w(\theta)/M) = \frac{\Pr(\theta \text{ and } u < w(\theta)/M)}{\Pr(u < w(\theta)/M)} = \frac{\Pr(u < w(\theta)/M \mid \theta)g(\theta)}{\Pr(u < w(\theta)/M)}.$$

- But,
  - $\Pr(u < w(\theta)/M \mid \theta) = w(\theta)/M$ since $u \sim U(0,1)$, and
  - $$\begin{aligned}\Pr(u < w(\theta)/M) &= \int \Pr(u < w(\theta)/M \mid \theta)g(\theta)d\theta \\ &= \int w(\theta)/Mg(\theta)d\theta = 1/M \int w(\theta)g(\theta)d\theta = 1/M \int p(\theta)d\theta = 1/M.\end{aligned}$$
- Therefore,

$$f(\theta|u < w(\theta)/M) = \frac{\Pr(u < w(\theta)/M \mid \theta)g(\theta)}{\Pr(u < w(\theta)/M)} = \frac{w(\theta)/Mg(\theta)}{1/M} = w(\theta)g(\theta) = p(\theta).$$

Rejection sampling for truncated densities

- The inverse CDF method works well for truncated densities but what happens when we can not write down the truncated CDF?
- Suppose we want to sample from $f_{[a,b]}(\theta)$, that is, a known pdf $f(\theta)$ truncated to $[a, b]$.
  - Recall that $f_{[a,b]}(\theta) \propto f(\theta)1[\theta \in [a, b]]$. Using the notation for rejection sampling, $p(\theta) = f_{[a,b]}(\theta)$ and $g(\theta) = f(\theta)$.
  - Set $1/M = \int_a^b f(\theta^\star)\mathrm{d}\theta^\star$, so that M is the normalizing constant of the truncated density.
  - Then, $w(\theta) = p(\theta)/g(\theta) = M1[\theta \in [a, b]] \leq M$ as required.

Rejection sampling for truncated densities

- We can then use the procedure on page 5 to generate the required samples.
- Specifically,
  - For each $i = 1, \ldots, m$, generate $\theta_i \sim f$. If $\theta_i \in [a, b]$, accept θ_i, otherwise **reject** and try again.
  - Easy to show that this is equivalent to accepting each θ_i with probability $w(\theta)/M$.

Example

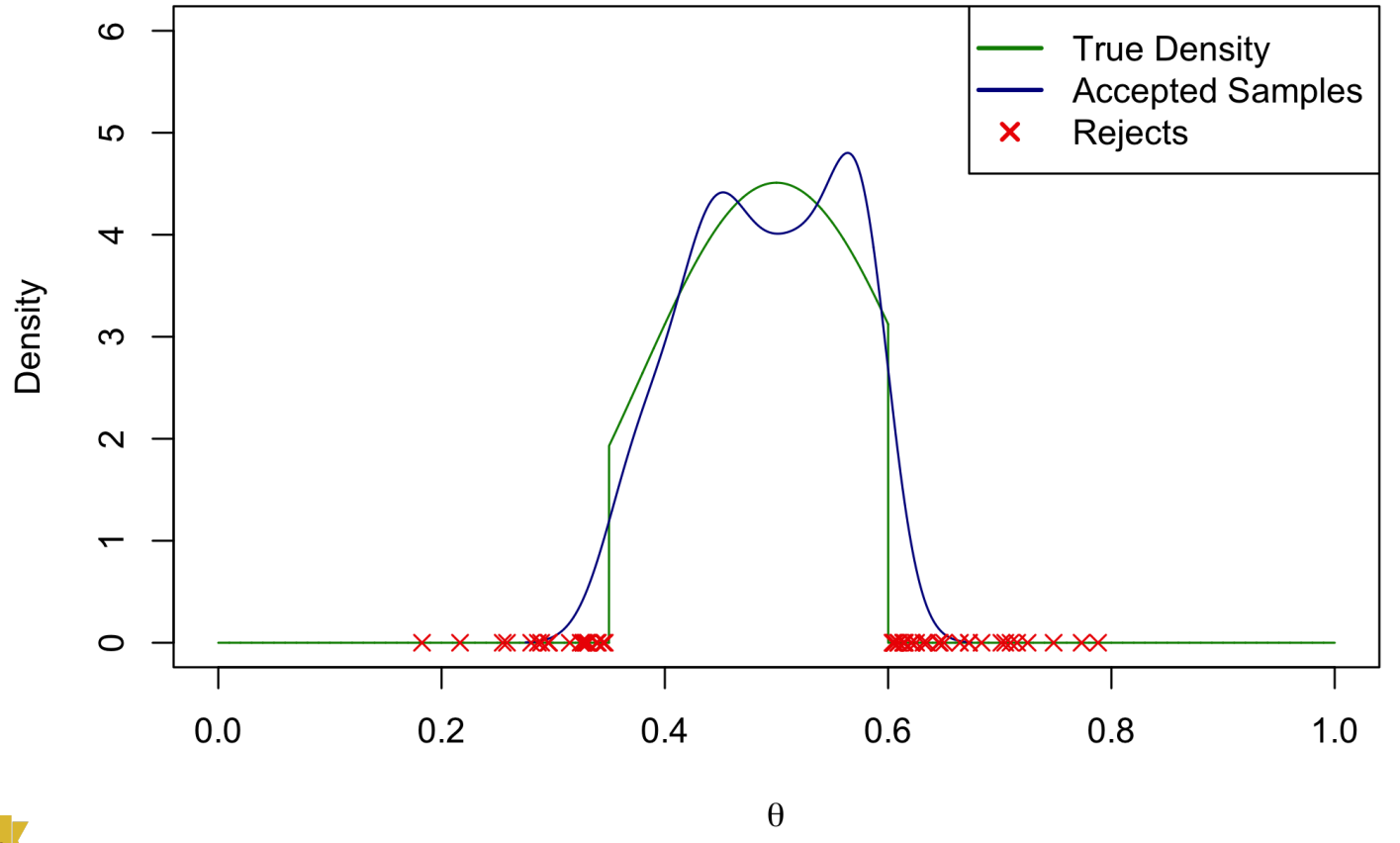
```
#Simple code for using rejection sampling to generate m samples
#from the Beta[10,10] density truncated to (0.35,0.6).
set.seed(12345)
#NOTE: there are more efficient ways to write this code!

#set sample size and reate vector to store sample
m <- 10000; THETA <- rep(0,m)
#keep track of rejects
TotalRejects <- 0; Rejections <- NULL
#now the 'for loop'
for(i in 1:m){
  t <- 0
  while(t < 1){
    theta <- rbeta(1,10,10)
    if(theta > 0.35 & theta < 0.6){
      THETA[i] <- theta
      t <- 1
    } else {
    TotalRejects <- TotalRejects + 1
    Rejections <- rbind(Rejections,theta)
  }
}
}
#How many rejections in all, to generate m=10000 samples?
TotalRejects
```

```
## [1] 3740
```

Acceptance rate ≈ 0.726.

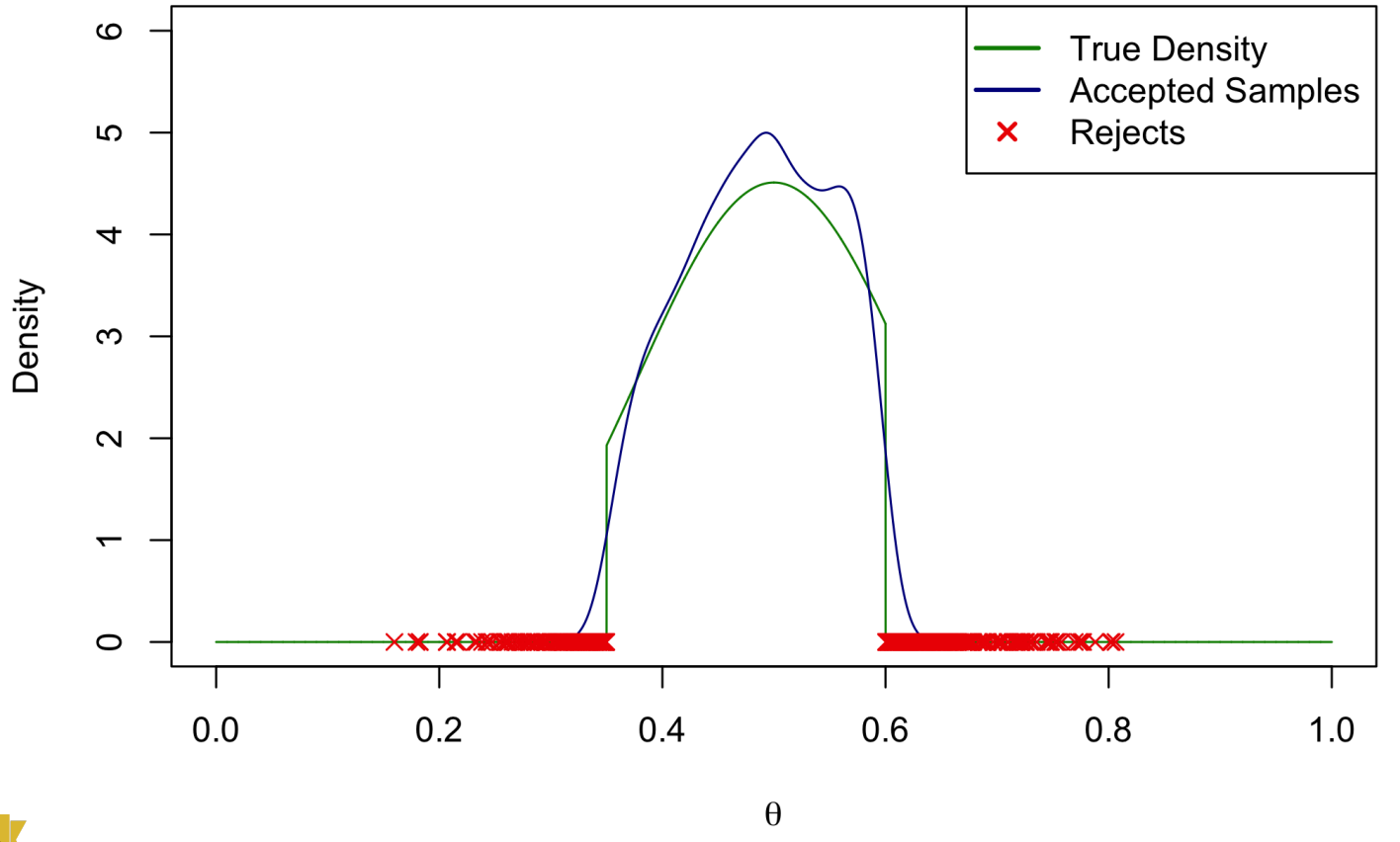
Example

How does our sample compare to the true truncated density? $m = 100$



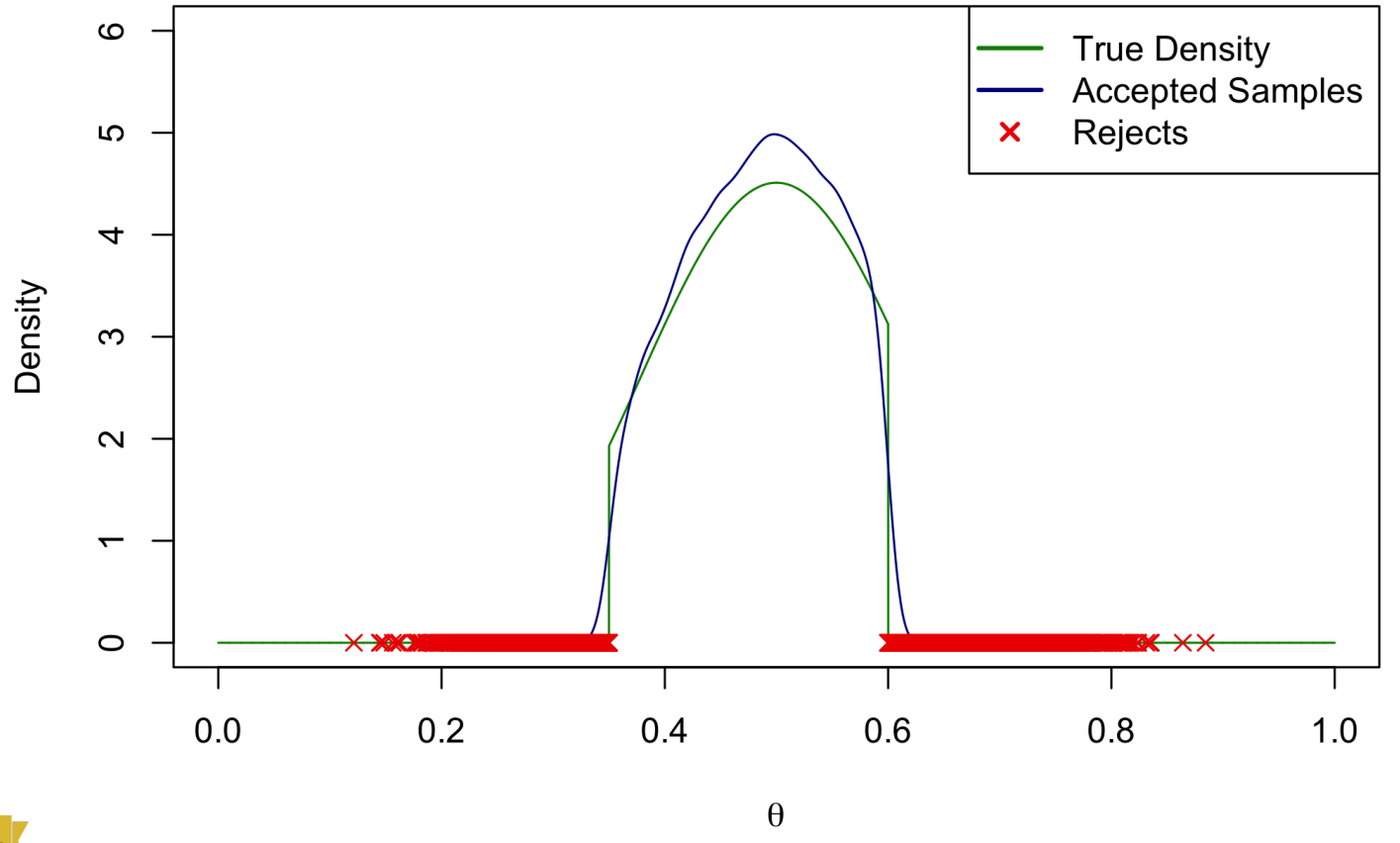
Example

How does our sample compare to the true truncated density? $m = 1000$



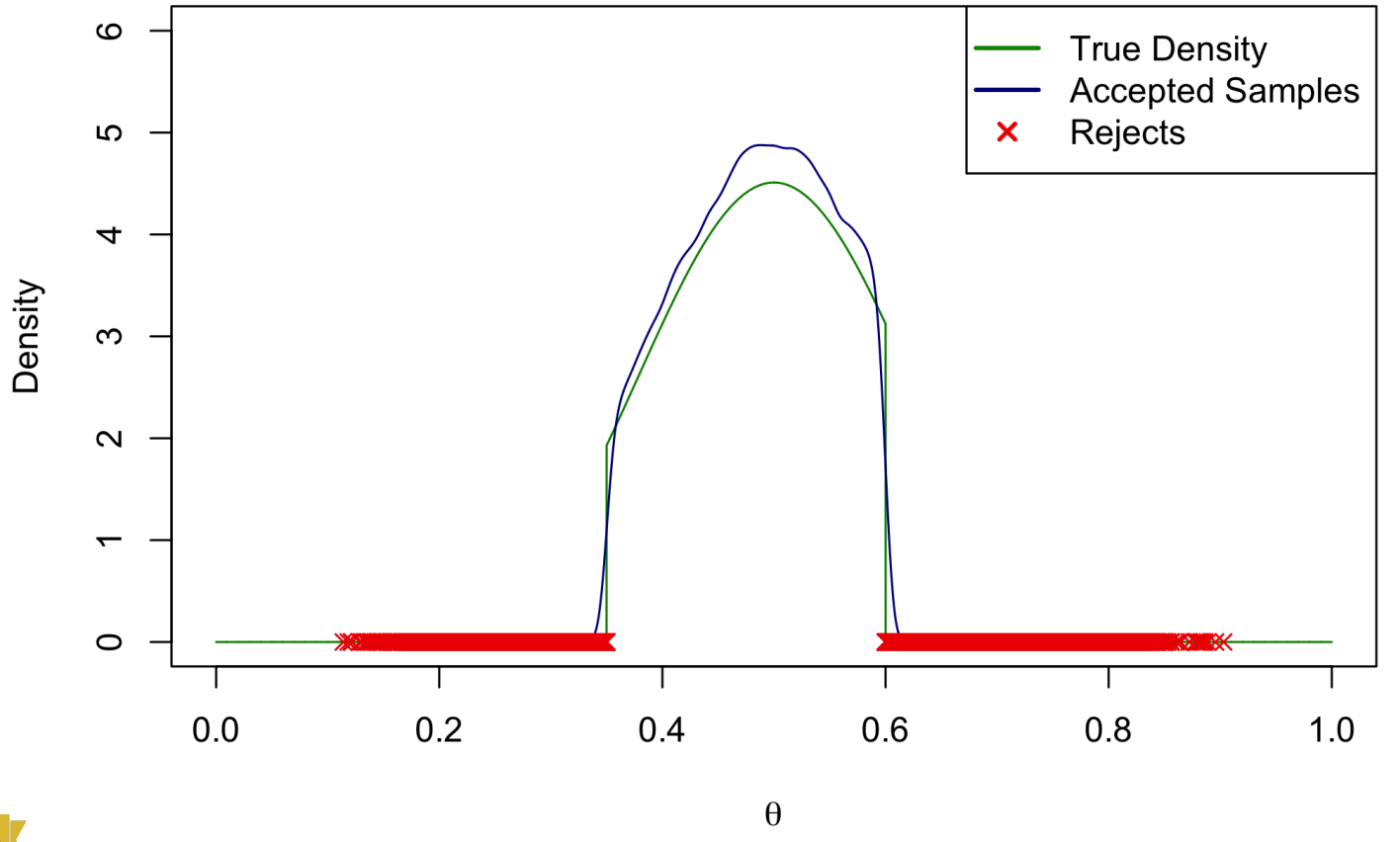
Example

How does our sample compare to the true truncated density? $m = 10000$



Example

How does our sample compare to the true truncated density? $m = 100000$



Comments

- Clearly less efficient than the inverse CDF method, which we already know how to use for this particular problem.
- When you can write down the truncated CDF, use the inverse CDF method instead.
- When you cannot, rejection sampling can be a possible alternative, as are many more sampling methods which we will not cover in this course.
- Anyway, generally, rejection sampling can still be very useful.
- Importance sampling is another related sampling method but we will not spend time on it. If you are interested, take a look at the next few slides. If not, feel free to skip to the normal model.

Importance sampling

- Importance sampling is actually one of the first steps into Monte Carlo analysis, in which simulated values from one distribution are used to explore another.
- Simulation from the "wrong distribution" can be incredibly useful as we have seen with rejection sampling and will also see later in this course.
- Not used as often anymore but still of practical interest in
  - fairly small problems, in terms of dimension,
  - in which the density of the distribution of interest can be easily evaluated, but when it is difficult to sample from directly, and
  - when it is relatively easy to identify and simulate from distributions that approximate the distribution of interest.
- Importance sampling and Rejection sampling use the same importance ratio ideas, but the latter leads to exact corrections and so exact samples from $p(\theta)$.

Importance sampling

- Interest lies in expectations of the form (instead of the actual samples)

$$H = \int h(\theta)p(\theta)d\theta,$$

- Write

$$H = \int h(\theta)w(\theta)g(\theta)d\theta \quad \text{with} \quad w(\theta) = p(\theta)/g(\theta)$$

that is, $\mathbb{E}[h(\theta)]$ under $p(\theta)$ is just $\mathbb{E}[h(\theta)w(\theta)]$ under $g(\theta)$.

- Using direct Monte Carlo integration

$$\bar{h} = \frac{1}{m}\sum_{i=1}^{m} w(\theta_i)h(\theta_i).$$

where $\theta_1, \ldots, \theta_m \overset{ind}{\sim} g(\theta)$. We are sampling from the "wrong" distribution.

Importance sampling

- The measure of "how wrong" we are at each simulated θ_m value is the importance weight

$$w(\theta_i) = p(\theta_i)/g(\theta_i).$$

  These ratios weight the sample estimates $h(\theta_i)$ to "correct" for the fact that we sampled the wrong distribution.

- See Lopes & Gamerman (Ch 3.4) and Robert and Casella (Ch. 3.3) for discussion of convergence and optimality.

- Clearly, the closer g is to p, the better the results, just as we had with rejection sampling.

Importance sampling

- Key considerations:
  - MC estimate $\bar{h}$ has the expectation H; and is generally almost surely convergent to H (under certain conditions of course but we will not dive into those).
  - $\mathbb{V}[\bar{h}]$ is often going to be finite in cases in which, generally, $w(\theta) = p(\theta)/g(\theta)$ is bounded and decays rapidly in the tails of $p(\theta)$.
  - Thus, superior MC approximations, are achieved for choices of $g(\theta)$ whose tails dominate those of the target $p(\theta)$.
  - That is, importance sampling distributions should be chosen to have tails at least as fat as the target (think normal distribution vs t-distribution).
  - Obviously require the support of $g(\theta)$ to be the same as, or contain, that of $p(\theta)$.
- These also clearly apply to rejection sampling too.

Importance sampling

- Problems in which $w(\theta) = p(\theta)/g(\theta)$ can be computed are actually rare.
- As you will see when we move away from conjugate distributions, we usually only know $p(\theta)$ up to a normalizing constant.
- When this is the case, simply "re-normalize" the importance weights, so that

$$\bar{h} = \frac{1}{m}\sum_{i=1}^{m} w_i h(\theta_i) \quad \text{where} \quad w_i = \frac{w(\theta_i)}{\sum_{i=1}^{m} w(\theta_i)}.$$

- Generally, in importance sampling, weights that are close to uniform are desirable, and very unevenly distributed weights are not.

Introduction to the univariate normal model

Motivating example: job training

- In the 1970s, researchers in the U.S. ran several randomized experiments intended to evaluate public policy programs.
- One of the most famous experiments is the National Supported Work (NSW) Demonstration, in which researchers wanted to assess whether or not job training for disadvantaged workers had an effect on their wages.
- Eligible workers were randomly assigned either to receive job training or not to receive job training.
- Candidates eligible for the NSW were randomized into the program between March 1975 and July 1977.
- For more details, read Lalonde, R. J. (1986) and Dehejia, R., and Wahba, S. (1999).

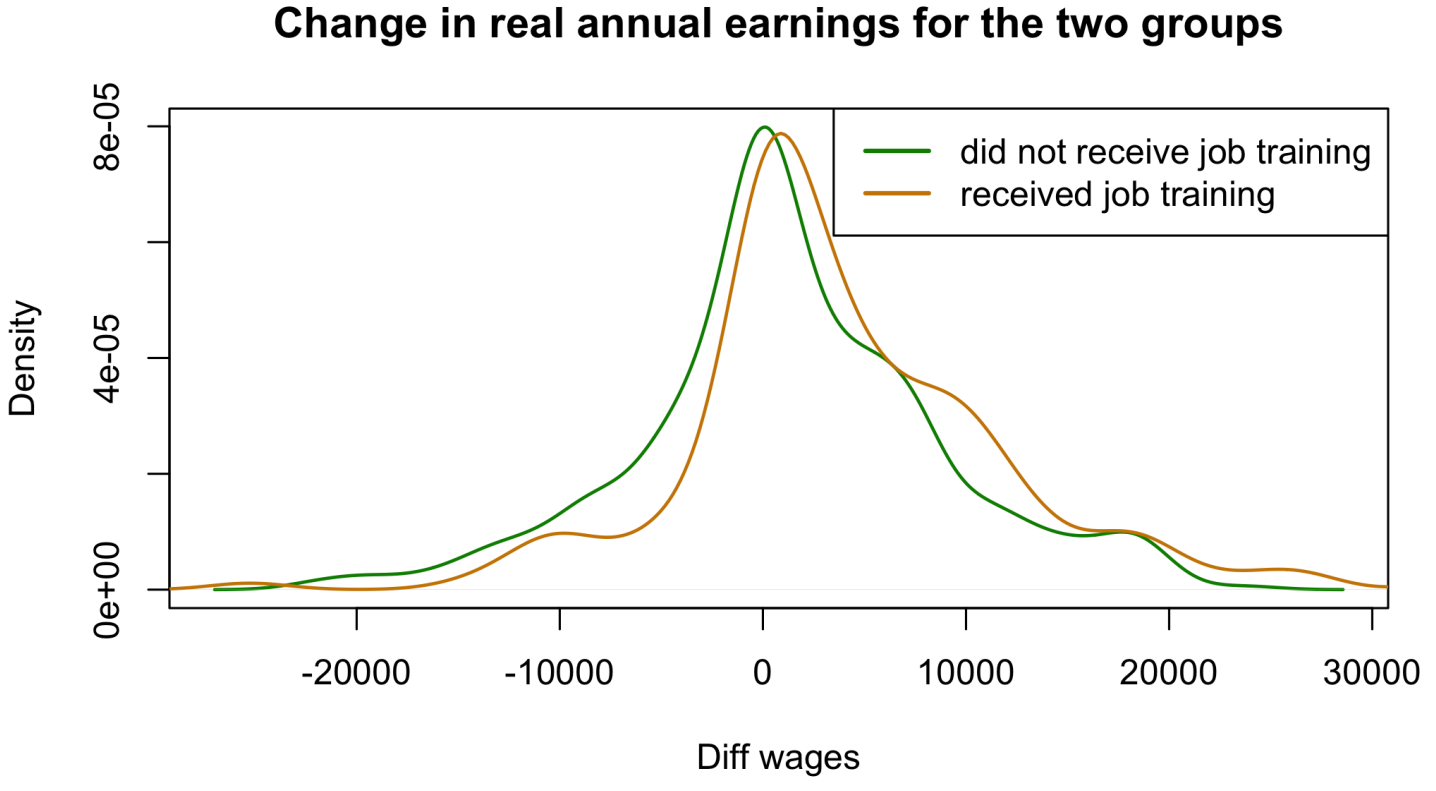
Motivating example: job training

- Setup:
  - **Pre-training wages**: real annual earnings in 1974 before training.
  - Two groups: some participants received job training and the rest did not.
  - **Post-training wages**: real annual earnings in 1978 upon completion of training.
- Question of interest: is there evidence that workers who receive job training tend to earn higher wages than workers who do not receive job training?
- The original study really is a causal inference setup, but the data used in this example only uses a subset of the data.
- The data is richer than what we will use it for (i.e., there are covariates we can control for) but we will only focus on the pre and post wages for the two groups.

JOB TRAINING: THE DATA

- Data:
  - No training group (N): sample size $n_N = 429$.
  - Training group (T): sample size $n_A = 185$.
  - **Diff wages**: Post-training wages – Pre-training wages.
- Summary statistics for change in annual earnings:
  - $\bar{y}_N = 1364.93$; $\sigma_N = 7460.05$
  - $\bar{y}_T = 4253.57$; $\sigma_T = 8926.99$
- Wages/income are well known to be approximately normally distributed. Let's look at the distribution of "change in annual earnings" for the two groups.

JOB TRAINING: THE DATA



Not completely normal but not too far off either. Lots of overlap between the two groups.

Model for changes in earnings

- $$y_i^{(T)} \sim \mathcal{N}(\mu_T, \sigma_T^2)$$

  $$y_i^{(N)} \sim \mathcal{N}(\mu_N, \sigma_N^2)$$

- Want posterior distribution of $\mu_T - \mu_N$. Specifically, we would like to compute $\Pr[\mu_T > \mu_N | Y_T, Y_N)$ or equivalently, $\Pr[\mu_T - \mu_N > 0 | Y_T, Y_N)$.
- Inference for $\mu_T - \mu_N$ can be complicated in frequentist paradigm when $\sigma_T^2 \neq \sigma_N^2$.
- Use approximate t-distributions based on the Welch-Satterthwaite degrees of freedom.
- Trivial with Bayesian inference
- By the way, also trivial to compute $\Pr[\sigma_T^2 > \sigma_N^2 | Y_T, Y_N)$ with Bayesian inference, which we will do later.
- How to do posterior inference for such normal models?

Another example: Pygmalion study

- Pygmalion effect is a phenomenon where expectation affects performance.
- Question of interest: do teachers' expectations impact academic development of children?
- Setup:
  - Researchers gave IQ test to elementary school children.
  - Randomly picked six children & told teachers that the test predicts them to **have high potential for accelerated growth**.
  - They randomly picked six children and told teachers that the test predicts them to have **NO potential for growth**.
  - At end of school year, they gave IQ test again to all students.
  - They recorded the change in IQ scores of each student.

Another example: Pygmalion study

- Data:
  - Accelerated group (A): 20, 10, 19, 15, 9, 18.
  - No growth group (N): 3, 2, 6, 10, 11, 5.
- Summary statistics:
  - $\bar{y}_A = 15.2$; $\sigma_A = 4.7$.
  - $\bar{y}_N = 6.2$; $\sigma_N = 3.6$.
- IQ test scores are also well known to be approximately normally distributed.
- Can't really check this assumption with only $n = 6$ observations.

Model for changes in scores

- $$y_i^{(A)} \sim \mathcal{N}(\mu_A, \sigma_A^2)$$

  $$y_i^{(N)} \sim \mathcal{N}(\mu_N, \sigma_N^2)$$

- Once again, we want posterior distribution of $\mu_A - \mu_N$.
- As before, we would like to compute $\Pr[\mu_A > \mu_N | Y_A, Y_N) \equiv \Pr[\mu_A - \mu_N > 0 | Y_A, Y_N)$.
- We would also like to compute $\Pr[\sigma_A^2 > \sigma_N^2 | Y_A, Y_N)$.
- To answer both questions, let's learn the Bayesian normal model.

Normal distribution

- A random variable Y has a normal distribution, written as $Y \sim \mathcal{N}(\mu, \sigma^2)$, if the pdf is

$$p(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}; \quad y \in (-\infty, \infty), \ \ \mu \in (-\infty, \infty), \ \ \sigma \in (0, \infty).$$

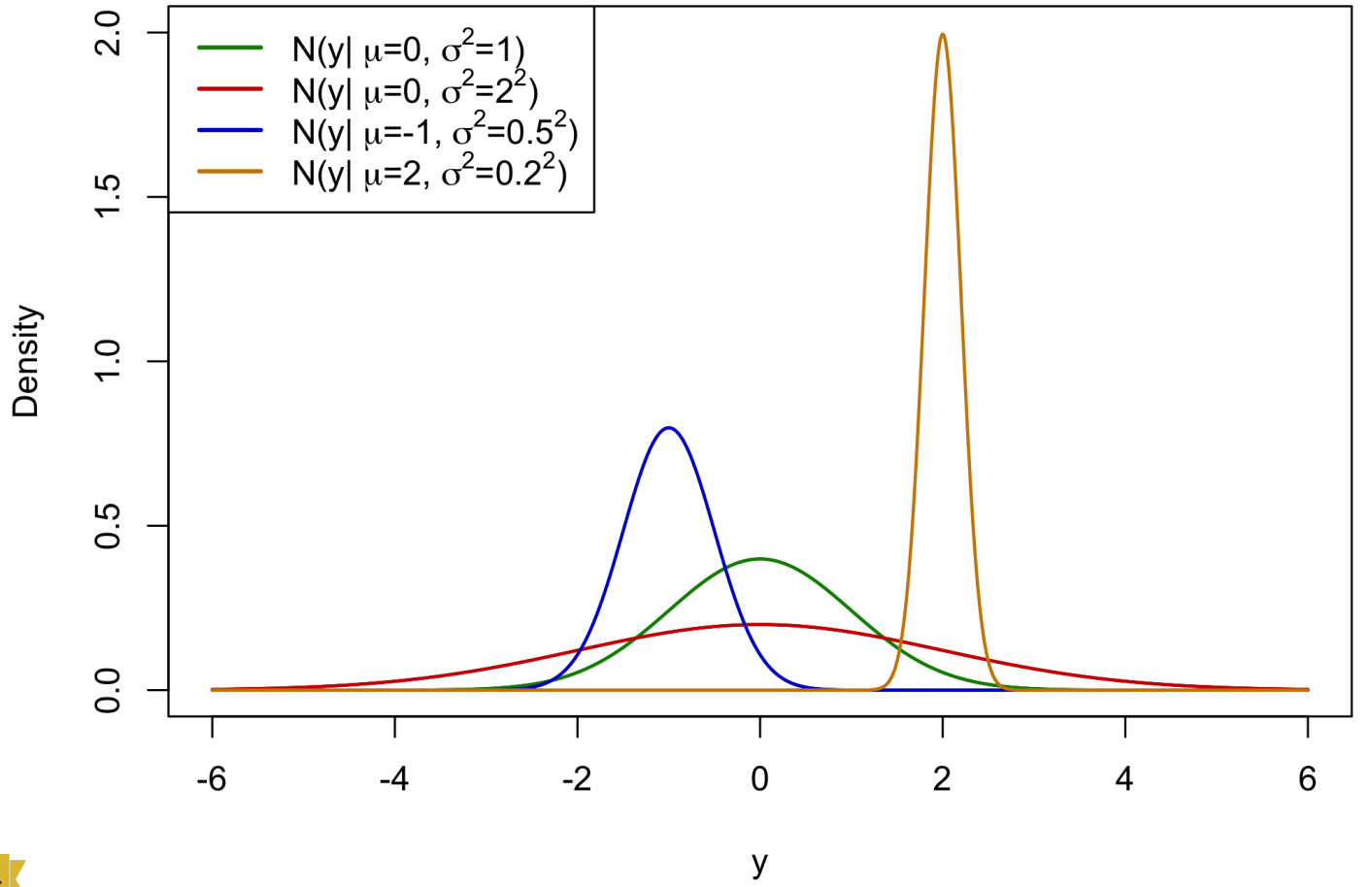
  where μ is the mean and σ^2 is the variance.

- It is also common (and would often be more convenient for our purposes) to write the pdf in terms of precision, τ, where $\tau = 1/\sigma^2$.

- In that case, the pdf is instead

$$p(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}} \tau^{\frac{1}{2}} e^{-\frac{1}{2}\tau(y-\mu)^2}; \quad y \in (-\infty, \infty), \ \ \mu \in (-\infty, \infty), \ \ \tau \in (0, \infty).$$

Example normal distributions



Comments on the normal distribution

- It is amazing how often real data are close to normally distributed.
- Likely a consequence of CLT – sums and means of independent random variables tend to be approximately normally distributed.
- Occurs under very general conditions.
- Normality?
  - Height, weight and other body measurements,
  - Income\wages\earnings,
  - Cumulative hydrologic measures such as annual rainfall or monthly river discharge,
  - Errors in astronomical or physical observations,
  - Many more examples!

Properties of the normal distribution

- Mean, median and mode are all the same (μ).
- Symmetric about the mean μ.
- 95% of the density (95% probability) within $\pm 1.96\sigma$ (approximately two standard deviations) of the mean.
- If $X \sim \mathcal{N}(\theta, s^2)$ and $Y \sim \mathcal{N}(\mu, \sigma^2)$ with $X \perp Y$, then

$$aX + bY \sim \mathcal{N}(a\theta + b\mu, a^2 s^2 + b^2 \sigma^2),$$

  for constants a and b.
- When independence does not hold, the sum of two normally distributed random variables is still normally distributed.
- However, when that is the case, we must account for the correlation in the variance term.

Notes on normal distribution in R

- `rnorm`, `dnorm`, `pnorm`, `qnorm` in R take mean and **standard deviation** σ as arguments.
- If you use the variance σ^2 instead you will get wrong answers!
- For example, `rnorm(n,m,s)` generates n normal random variables with mean m and standard deviation s, that is, $\mathcal{N}(m, s^2)$.

Normal model

- Suppose we have independent observations $Y = (y_1, y_2, \ldots, y_n)$, where each $y_i \sim \mathcal{N}(\mu, \sigma^2)$ or $y_i \sim \mathcal{N}(\mu, \tau^{-1})$, with unknown parameters μ and σ^2 (or τ).
- Then, the likelihood is

$$
\begin{aligned}
L(Y; \mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \tau^{\frac{1}{2}} \exp\left\{ -\frac{1}{2}\tau(y_i - \mu)^2 \right\} \\
&\propto \tau^{\frac{n}{2}} \exp\left\{ -\frac{1}{2}\tau \sum_{i=1}^{n} (y_i - \mu)^2 \right\} \\
&\propto \tau^{\frac{n}{2}} \exp\left\{ -\frac{1}{2}\tau \sum_{i=1}^{n} \left[(y_i - \bar{y}) - (\mu - \bar{y}) \right]^2 \right\} \\
&\propto \tau^{\frac{n}{2}} \exp\left\{ -\frac{1}{2}\tau \left[\sum_{i=1}^{n} (y_i - \bar{y})^2 - \sum_{i=1}^{n} (\mu - \bar{y})^2 \right] \right\} \\
&\propto \tau^{\frac{n}{2}} \exp\left\{ -\frac{1}{2}\tau \left[\sum_{i=1}^{n} (y_i - \bar{y})^2 - n(\mu - \bar{y})^2 \right] \right\} \\
&\propto \tau^{\frac{n}{2}} \exp\left\{ -\frac{1}{2}\tau s^2 (n-1) \right\} \exp\left\{ -\frac{1}{2}\tau n (\mu - \bar{y})^2 \right\}.
\end{aligned}
$$

Likelihood for normal model

- Likelihood:

$$L(Y;\mu,\sigma^2) \propto \tau^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\tau s^2(n-1)\right\} \exp\left\{-\frac{1}{2}\tau n(\mu-\bar{y})^2\right\},$$

  where

  - $\bar{y} = \sum_{i=1}^n y_i$ is the sample mean; and
  - $s^2 = \sum_{i=1}^n (y_i - \bar{y})^2/(n-1)$ is the sample variance.

- Sufficient statistics:

  - Sample mean $\bar{y}$; and
  - Sample sum of squares $SS = s^2(n-1) = \sum_{i=1}^n (y_i - \bar{y})^2$.

- MLEs:

  - $\hat{\mu} = \bar{y}$.
  - $\hat{\tau} = n/SS$, and $\hat{\sigma}^2 = SS/n$.

Inference for mean, conditional on variance

- We can break down inference problem for this two-parameter model into two one-parameter problems.
- First start by developing inference on μ when σ^2 is known. Turns out we can use a conjugate prior for $\pi(\mu|\sigma^2)$. We will get to unknown σ^2 in the next class.
- For σ^2 known, the normal likelihood further simplifies to

$$\propto \exp\left\{-\frac{1}{2}\tau n(\mu-\bar{y})^2\right\},$$

  leaving out everything else that does not depend on μ.
- For $\pi(\mu|\sigma^2)$, we consider $\mathcal{N}(\mu_0, \sigma_0^2)$, i.e., $\mathcal{N}(\mu_0, \tau_0^{-1})$, where $\tau_0^{-1} = \sigma_0^2$.
- Let's derive the posterior $\pi(\mu|Y, \sigma^2)$.

Inference for mean, conditional on variance

- Posterior:

$$\pi(\mu|Y,\sigma^2) \propto \pi(\mu|\sigma^2)L(Y;\mu,\sigma^2) \propto \exp\left\{-\frac{1}{2}\tau_0(\mu-\mu_0)^2\right\} \exp\left\{-\frac{1}{2}\tau n(\mu-\bar{y})^2\right\}$$

- Expanding out squared terms

$$\Rightarrow \pi(\mu|Y,\sigma^2) \propto \exp\left\{-\frac{1}{2}\tau_0(\mu^2-2\mu\mu_0+\mu_0^2)\right\} \exp\left\{-\frac{1}{2}\tau n(\mu^2-2\mu\bar{y}+\bar{y}^2)\right\}$$

- Ignoring terms not containing μ

$$\begin{aligned}\Rightarrow \pi(\mu|Y,\sigma^2) &\propto \exp\left\{-\frac{1}{2}\tau_0(\mu^2-2\mu\mu_0)\right\} \exp\left\{-\frac{1}{2}\tau n(\mu^2-2\mu\bar{y})\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[\tau_0(\mu^2-2\mu\mu_0)+\tau n(\mu^2-2\mu\bar{y})\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[\mu^2(\tau n+\tau_0)-2\mu(\tau n\bar{y}+\tau_0\mu_0)\right]\right\}.\end{aligned}$$

Inference for mean, conditional on variance

- Notice that $\left[\mu^2(\tau n + \tau_0) - 2\mu(\tau n \bar{y} + \tau_0 \mu_0)\right]$ is essentially a quadratic equation of the form $a^\star \mu^2 - 2b^\star \mu + c^\star$, where
  - $a^\star = \tau n + \tau_0$,
  - $b^\star = \tau n \bar{y} + \tau_0 \mu_0$, and
  - $c^\star$ does not depend on μ.

    Note that $c^\star$ contains some of the terms we ignored on the previous slide but we need not know its exact form here.
- Goal: Turn this quadratic equation into an expression of the form $m(\mu - r)^2$, for some m and r, so that we have something that resembles the kernel of a normal density.
- How? Complete the square!

Inference for mean, conditional on variance

- Recall how to complete the square. Specifically, we can write

$$a\mu^2 + b\mu + c$$

as

$$a(\mu + d)^2 + e,$$

where

- $d = \dfrac{b}{2a}$, and
- $e = c - \dfrac{b^2}{4a}$.

Inference for mean, conditional on variance

- Completing the square and rearranging (where $a^\star = \tau n + \tau_0$ and $b^\star = \tau n \bar{y} + \tau_0 \mu_0$),

$$
\begin{aligned}
\Rightarrow \pi(\mu | Y, \sigma^2) &\propto \exp\left\{-\frac{1}{2}\left[a^\star \mu^2 - 2b^\star \mu\right]\right\} \\
&= \exp\left\{-\frac{1}{2} a^\star \left[\mu^2 - \frac{2b^\star}{a^\star}\mu\right]\right\} \\
&= \exp\left\{-\frac{1}{2} a^\star \left[\mu^2 - \frac{2b^\star}{a^\star}\mu + \frac{(b^\star)^2}{(a^\star)^2}\right] + \frac{(b^\star)^2}{2a^\star}\right\} \\
&\propto \exp\left\{-\frac{1}{2} a^\star \left[\mu^2 - \frac{b^\star}{a^\star}\right]^2\right\},
\end{aligned}
$$

which is the kernel of a normal distribution with

- mean $\frac{b^\star}{a^\star}$, and
- precision $a^\star$ or variance $(a^\star)^{-1}$.

Posterior with precision terms

- In terms of precision, we have

$$\mu|Y,\sigma^2 \sim \mathcal{N}(\mu_n, \tau_n^{-1})$$

where

$$\mu_n = \frac{b^\star}{a^\star} = \frac{\tau n \bar{y} + \tau_0 \mu_0}{\tau n + \tau_0}$$

and

$$\tau_n = a^\star = \tau n + \tau_0.$$

Posterior with precision terms

- As mentioned before, Bayesians often prefer to talk about precision instead of variance.
- We have
  - τ as the sampling precision (how close the y_i's are to μ).
  - τ_0 as the prior precision (our prior belief about the uncertainty about μ around our prior guess μ_0).
  - τ_n as the posterior precision
- As we have on the previous slide, *the posterior precision equals the prior precision plus the data precision*.
- That is, we see that the posterior information is a sum of the prior information and the information from the data.

Posterior with precision terms: combining information

- Posterior mean is weighted sum of prior information plus data information:

$$\begin{aligned}\mu_n &= \frac{n\tau\bar{y} + \tau_0\mu_0}{\tau n + \tau_0} \\ &= \frac{\tau_0}{\tau_0 + \tau n}\mu_0 + \frac{n\tau}{\tau_0 + \tau n}\bar{y}\end{aligned}$$

- Recall that σ^2 (and thus τ) is known for now.
- If we think of the prior mean as being based on κ_0 prior observations from a similar population as $y_1, y_2, \ldots, y_n$, then we might set $\sigma_0^2 = \frac{\sigma^2}{\kappa_0}$, which implies $\tau_0 = \kappa_0\tau$, and then the posterior mean is given by

$$\mu_n = \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{y}.$$

Posterior with variance terms

- In terms of variances, we have

$$\mu|Y,\sigma^2 \sim \mathcal{N}(\mu_n, \sigma_n^2)$$

where

$$\mu_n = \frac{b^\star}{a^\star} = \frac{\frac{n}{\sigma^2}\bar{y} + \frac{1}{\sigma_0^2}\mu_0}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}$$

and

$$\sigma_n^2 = \frac{1}{a^\star} = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}.$$

- It is still easy to see that we can re-express the posterior information as a sum of the prior information and the information from the data.

Noninformative and improper priors

- Clearly, we need to specify both μ_0 and τ_0 to do inference here.
- When prior distributions have no population basis, that is, there is no justification of the prior as "prior data", prior distributions can be difficult to construct.
- To that end, there is often the desire to construct noninformative priors, with the rationale being *"to let the data speak for themselves"*.
- For example, we could instead assume a uniform prior on μ that is constant over the real line, i.e., $\pi(\mu) \propto 1 \Rightarrow$ all values on the real line are equally likely apriori.
- Clearly, this is not a valid pdf since it will not integrate to 1 over the real line. Such priors are known as improper priors.
- An improper prior can still be very useful, we just need to ensure it results in a proper posterior.

Jeffreys' prior

- Question: is there a prior pdf (for a given model) that would be universally accepted as a noninformative prior?
- Laplace proposed the uniform distribution. His proposal is not universally accepted because it lacks invariance under monotone transformations of the parameter.
- For example, a uniform prior on the binomial proportion parameter θ is not the same as a uniform prior on the odds parameter $\phi = \frac{\theta}{1-\theta}$, which is not ideal.
- A more acceptable approach was introduced by Jeffreys. For single parameter models, the Jeffreys' prior defines a noninformative prior density of a parameter θ as

$$\pi(\theta) \propto \sqrt{\mathcal{I}(\theta)}$$

where $\mathcal{I}(\theta)$ is the Fisher information for θ.

Jeffreys' prior

- The Fisher information gives a way to measure the amount of information a random variable Y carries about an unknown parameter θ of a distribution that describes Y.
- Formally, $\mathcal{I}(\theta)$ is defined as

$$\mathcal{I}(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(y;\theta)\right)^2 \middle| \theta\right] = \int_y \left(\frac{\partial}{\partial\theta}\log f(y;\theta)\right)^2 f(y;\theta)dy.$$

- Alternatively,

$$\mathcal{I}(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial^2\theta}\log f(y;\theta)\middle|\theta\right] = \int_y \left(\frac{\partial^2}{\partial^2\theta}\log f(y;\theta)\right) f(y;\theta)dy.$$

- Turns out that the Jeffreys' prior for μ under the normal model, when σ^2 is known, is

$$\pi(\mu) \propto 1,$$

the uniform prior over the real line. You should try to derive this.

Inference for mean, conditional on variance using Jeffreys' prior

- Recall that for σ^2 known, the normal likelihood simplifies to

$$\propto \exp\left\{-\frac{1}{2}\tau n(\mu-\bar{y})^2\right\},$$

  ignoring everything else that does not depend on μ.

- With the Jeffreys' prior $\pi(\mu) \propto 1$, can we derive the posterior distribution?

Inference for mean, conditional on variance using Jeffreys' prior

- Posterior:

$$\begin{aligned}\pi(\mu|Y,\sigma^2) &\propto \exp\left\{-\frac{1}{2}\tau n(\mu-\bar{y})^2\right\}\pi(\mu)\\ &\propto \exp\left\{-\frac{1}{2}\tau n(\mu-\bar{y})^2\right\}.\end{aligned}$$

- This is the kernel of a normal distribution with
  - mean $\bar{y}$, and
  - precision $n\tau$ or variance $\frac{1}{n\tau} = \frac{\sigma^2}{n}$.
- Written differently, we have $\mu|Y,\sigma^2 \sim \mathcal{N}(\bar{y}, \frac{\sigma^2}{n})$
- This should look familiar to you. Does it?
- To get back to our example, we need to extend inference to unknown σ^2 . We'll start there in the next class.