```r
knitr::opts_chunk$set(echo = FALSE, fig.path = '', cache = TRUE)
```

Part 1

Let $Y_1, Y_2, \ldots, Y_n|\mu, \lambda \overset{iid}{\sim} N(\mu, \lambda^{-1})$, and **assume that $\lambda > 0$ is fixed and known**.

1. Derive the exponential family form for the generating distribution (the likelihood) of $Y_1, Y_2, \ldots, Y_n$.

$$
\begin{aligned}
p(y_{1:n}|\mu) &= \prod_{i=1}^{n} \mathbf{1}_{\{y_i \in \mathbb{R}\}} (\frac{\lambda}{2\pi})^{\frac{1}{2}} \exp[-\frac{\lambda}{2}(y_i - \mu)^2] \\
&= \prod_{i=1}^{n} \mathbf{1}_{\{y_i \in \mathbb{R}\}} (\frac{\lambda}{2\pi})^{\frac{n}{2}} \exp[-\frac{\lambda}{2}\sum_{i=1}^{n}(y_i - \mu)^2] \\
&= \prod_{i=1}^{n} \mathbf{1}_{\{y_i \in \mathbb{R}\}} (\frac{\lambda}{2\pi})^{\frac{n}{2}} \exp[-\frac{\lambda}{2}(\sum_{i=1}^{n} y_i^2 - 2\mu\sum_{i=1}^{n} y_i + n\mu^2)] \\
&= \underbrace{\prod_{i=1}^{n} \mathbf{1}_{\{y_i \in \mathbb{R}\}} (\frac{\lambda}{2\pi})^{\frac{n}{2}} \exp[-\frac{\lambda}{2}\sum_{i=1}^{n} y_i^2]}_{h(y)} \exp[\underbrace{n\lambda\mu}_{\phi(\mu)} \underbrace{\bar{y}}_{t(y)} - n\underbrace{\frac{\lambda\mu^2}{2}}_{\kappa(\mu)}] \\
&\propto \exp[-\frac{n\lambda}{2}(\mu^2 - 2\mu\bar{y})]
\end{aligned}
$$

2. Use the likelihood derived above in exponential family form to show that the conjugate family of priors for $\boldsymbol{\mu}$ corresponds to $\left\{N(\mu_0, (\nu\lambda)^{-1}) : \mu_0 \in \mathbb{R}, \nu > 0\right\}$,

$$
p(\mu) = \mathbf{1}_{\{\mu \in \mathbb{R}\}} (\frac{\nu\lambda}{2\pi})^{\frac{1}{2}} \exp[-\frac{\nu\lambda}{2}(\mu - \mu_0)^2] \propto \exp[-\frac{\nu\lambda}{2}(\mu^2 - 2\mu\mu_0)]
$$

3. Use the results from the two parts above to find the posterior distribution for μ.

$$
\begin{aligned}
p(\mu|y_{1:n}) &\propto p(y_{1:n}|\mu, \lambda)p(\mu) \\
&\propto \exp[-\frac{n\lambda}{2}(\mu^2 - 2\mu\bar{y})] \exp[-\frac{\nu\lambda}{2}(\mu^2 - 2\mu\mu_0)] \\
&\propto \exp[-\frac{\lambda}{2}\left(\mu^2(n+\nu) - 2\mu(n\bar{y} + \nu\mu_0)\right)] \\
&\propto \exp[-\frac{\lambda(n+\nu)}{2}\left(\mu^2 - 2\mu\frac{n\bar{y} + \nu\mu_0}{n+\nu} + [\frac{n\bar{y} + \nu\mu_0}{n+\nu}]^2\right)] \exp[\frac{\lambda(n+\nu)}{2}(\frac{n\bar{y} + \nu\mu_0}{n+\nu})^2] \\
&\propto \exp[-\frac{\lambda(n+\nu)}{2}[\mu - (\frac{n}{n+\nu}\bar{y} + \frac{\nu}{n+\nu}\mu_0)]^2]
\end{aligned}
$$

which corresponds to the kernel of a $N(\mu^\star, \lambda^{\star -1})$, with precision $\lambda^\star = \lambda(n+\nu)$, and mean $\mu^\star = \frac{n}{n+\nu}\bar{y} + \frac{\nu}{n+\nu}\mu_0$

And so $\boldsymbol{\mu}|y_{1:n} \sim N(\mu^\star, \lambda^{\star -1})$, which implies that the normal family of priors for $\boldsymbol{\mu}$ is conjugate with the normal likelihood.

4. Now, assume that λ is not known, and show that the collection of $N(\mu, \lambda^{-1})$ distributions, with $\mu \in \mathbb{R}$ and $\lambda > 0$, is a two-parameter exponential family, and identify the sufficient statistics function $t(y_{1:n}) = (t_1(y_{1:n}), t_2(y_{1:n}))^T$ for your parametrization.

$$
\begin{aligned}
p(y_{1:n}|\mu,\lambda) &= \prod_{i=1}^{n} \mathbf{1}_{\{y_i\in\mathbb{R}\}}(\frac{\lambda}{2\pi})^{\frac{1}{2}}\exp[-\frac{\lambda}{2}(y_i-\mu)^2] \\
&= \underbrace{\mathbf{1}_{\{y_i\in\mathbb{R}\}}(2\pi)^{-\frac{n}{2}}}_{h(y)}\exp[\underbrace{-\frac{\lambda}{2}\sum_{i=1}^{n}y_i^2+n\lambda\mu\bar{y}}_{\phi(\mu,\lambda)t(y_{1:n})}-n\underbrace{\frac{\lambda\mu^2-\ln\lambda}{2}}_{\kappa(\mu)}]
\end{aligned}
$$

where $\phi(\mu,\lambda)=(-\frac{\lambda}{2},n\lambda\mu)$, $t(y_{1:n})=(\sum_{i=1}^{n}y_i^2,\bar{y})^T$

Part 2

Two competitors for the snowiest city in the world are Aomori City in Japan, and Valdez in the state of Alaska. Here are annual snowfall records, in inches/year, for the two cities. Assume that for each city independently, the data is i.i.d. normal.

5. Do you think an iid normal model is appropriate here? Why or why not?

These data are weather observations of many years in two locations. Before analys the data, Our prior experiences tell me that the atmospheric and geographic factors decide the precipitation.

- Independent:

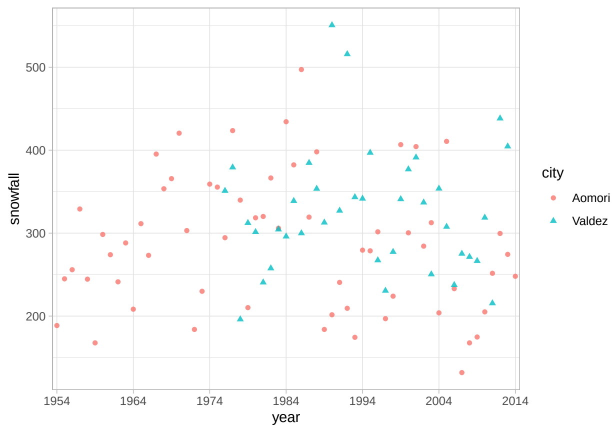
Not strictly. From a global perspective, all the climate issues are connected and have interaction with each other. Therefore, compared to a data in a lab, the weather observations are not purely independent. If the researcher think the distance between the two cities is far enough that the interations are negligible, we can consider the data as independent. The effect of time factor may not be independent, previous years value may affects the next year, but it doesn't affect the comparison of two cities.

- Identical:

Not sure. If the geographic factors are principal causes of weather, the distribution of snowfall in two cities may not be same. If the atomspheric factors are principal causes, the two cities' snowfall records happened in a same "rule" and are identical.

- Normality:

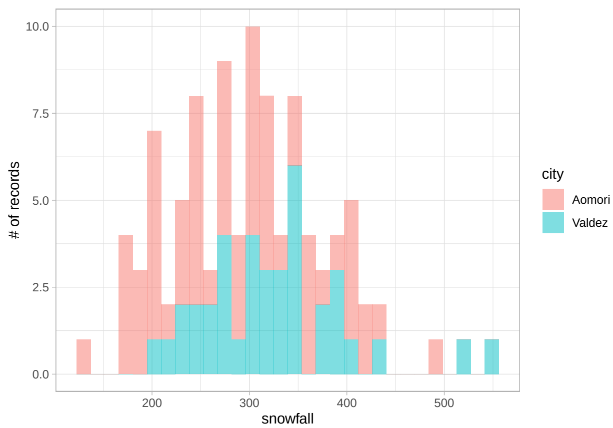
Not sure. We can say that there is a large number of facotrs on snowfall. If the effects of these factors are additive, the data may be normal distributed. From the central limit theorem, the sum or mean of a set of random variables is approximately normally distributed.



The figure of snowfall in the different years didn't show obvious trend over years

6. Is the mean annual snowfall for Valdez higher than that of Aomori? To address this question, perform an analysis like the one for the Pygmalion effect in the class notes assuming normality. In particular, your analysis should involve computing the posterior probability that the mean annual snowfall for Valdez is higher than that of Aomori. Choose suitable prior parameters (hyperparameters) that reflect your prior knowledge (or lack thereof).

```
## city: Aomori
##    n     mean      sd
## 1 61 286.9148 81.0554
## --------------------------------------------------------
## city: Valdez
##    n     mean       sd
## 1 38 325.3974 75.10118
```



The mean annual snowfall for Valdez(325.4 inches/year) is higher than that of Aomori (286.9 inches/year). The histograms show that the mean for Valdez was higher. The distributions do not look perfectly normal. The problem setup under the normality assumption is

$$\text{Valdez: } X_1, \dots, X_{n_v} \overset{iid}{\sim} N(\mu_v, \lambda_v^{-1})$$

$$\text{Aomori: } Y_1, \dots, Y_{n_a} \overset{iid}{\sim} N(\mu_a, \lambda_a^{-1})$$

We are interested in the difference between the means is $\mu_v > \mu_a$. We don't know the standard deviations $\sigma_v = \lambda_v^{-1/2}$ and $\sigma_a = \lambda_a^{-1/2}$. When $\sigma_v \neq \sigma_a$, The frequentist approach to this problem involves approximate t-distributions based on the Welch–Satterthwaite degrees of freedom. From the Bayesian perspective we need to determine is $P(\vec{\mu}_p > \vec{\mu}_c|\text{data})$

Assuming the parameters for the two citiess have the same distribution and are independent across cities. Their conjugate prior for the means and variances are given by

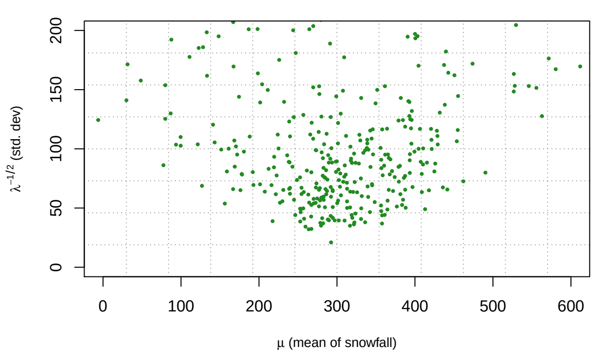
$$\text{Valdez: } \vec{\mu}_v, \vec{\lambda}_v \sim \text{Normal-Gamma}(\mu_0, \nu, \alpha, \beta)$$
$$\text{Aomori: } \vec{\mu}_a, \vec{\lambda}_a \sim \text{Normal-Gamma}(\mu_0, \nu, \alpha, \beta).$$

We use subjective prior knowledge to assign values for the parameters in the prior (a.k.a. hyperparameters), as

- $\mu_0 = 300$: we don't know the snowfall value on average.
- $\nu = 1$: we are uncertain about how big the mean will be, setting it to 1 is equivalent to assigning the strength of one observation to the information coming from the prior on μ_p and μ_c.
- $\alpha = 1/2$: refelcts uncertainty about how much the standard deviation is
- $\beta = 6400\alpha$ this value leads to a expected standard deviation of the changes in the snowfall of about 80 inches/year, since $\sqrt{E(\lambda)} = \sqrt{\alpha/\beta} = 1/80$, and so $\sigma \approx 80$ a priori.

Check whether a prior conforms to our beliefs?

- Draw some samples from the prior and look at them—this is probably the best general strategy.
- I choose to look at sample hypothetical datasets $x_{1:n_v}, y_{1:n_a}$ drawn using these sampled parameter values.
- Plot the cdf and check various quantiles (first quartile, median, third quartile), if univariate.
- Plot the pdf.
- Look at various moments (e.g., mean, standard deviation).

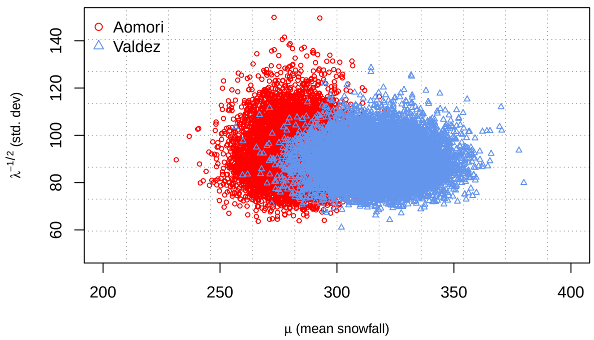


From the derivations in the previous section we have that the posterior distributions for $\vec{\mu}_v, \vec{\lambda}_v$ and for $\vec{\mu}_a, \vec{\lambda}_a$ are given by

$$\text{Valdez: } \vec{\mu}_v, \vec{\lambda}_v \sim \text{Normal-Gamma}(\mu_v^\star, \nu_v^\star, \alpha_v^\star, \beta_v^\star)$$
$$\text{Aomori: } \vec{\mu}_a, \vec{\lambda}_a \sim \text{Normal-Gamma}(\mu_a^\star, \nu_a^\star, \alpha_a^\star, \beta_a^\star).$$

where

Valdez		Aomori	
$\mu_v^\star = \frac{n_v}{\nu+n_v}\bar{x} + \frac{\nu}{\nu+n_v}\mu_0$	$=317.05$	$\mu_a^\star = \frac{n_a}{\nu+n_a}\bar{y} + \frac{\nu}{\nu+n_a}\mu_0$	$=282.29$
$\nu_v^\star = \nu + n_v$	$=39$	$\nu_a^\star = \nu + n_a$	$=62$
$\alpha_v^\star = \alpha + n_v/2$	$=19.5$	$\alpha_a^\star = \alpha + n_a/2$	$=31$
$\beta_v^\star = \frac{1}{2}(\sum_i x_i^2 + \nu\mu_0^2 + 2\beta - \nu_v^\star {\mu_v^\star}^2)$	$=1.591277 \times 10^5$	$\beta_a^\star = \frac{1}{2}(\sum_i y_i^2 + \nu\mu_0^2 + 2\beta - \nu_a^\star {\mu_a^\star}^2)$	$=2.407955 \times 10^5$



The Figure shows a scatterplot of samples from the posteriors by a Monte Carlo approximation: taking a bunch of samples (nsim$= 10^4$) from each of the posteriors and the fraction of times we have $\mu_v > \mu_a$

Valdez: $\vec{\mu}_v, \vec{\lambda}_v \sim$ Normal-Gamma(317.1,39,19.5,1.591277×10^5)

Aomori: $\vec{\mu}_a, \vec{\lambda}_a \sim$ Normal-Gamma(282.3,62,31,2.407955×10^5)

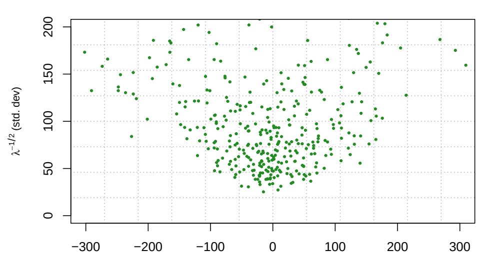
and the approximate $P(\mu_v > \mu_a | x_{1:n_p}, y_{1:n_c}) \approx \frac{1}{\text{nsim}} \sum_{k=1}^{\text{nsim}} \mathbf{1}_{\{\mu_v^{(k)} > \mu_a^{(k)}\}}$=0.967

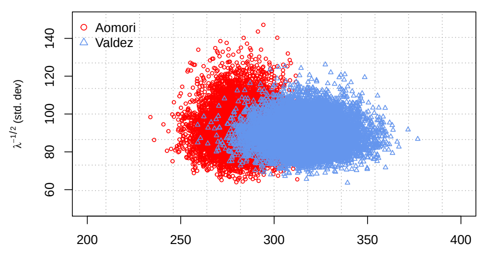
This data strongly supports the hypothesis that Valdez has more snowfll. This is evidenced in both the figure and by the probability calculated, the posterior probability that the Valdez has a higher mean of snowfall is about 0.967.

7. Try different values for the hyperparameters, to see what effect they have on the results. Report your results for three different settings of the hyperparameters.

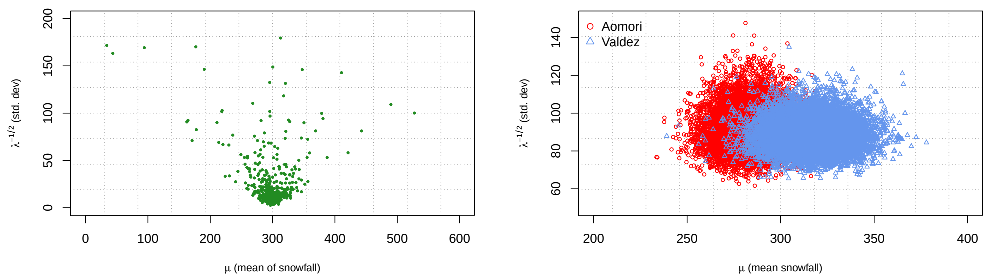
I tried four combination of μ_0, nu and β. I did not try to change the value of α without obvious evidents,

	μ_0	nu	α	β
set 1	0	1	1/2	6400α
set 2	300	1	1/2	100α
set 3	300	4	1/2	6400α
set 4	300	1/4	1/2	6400α

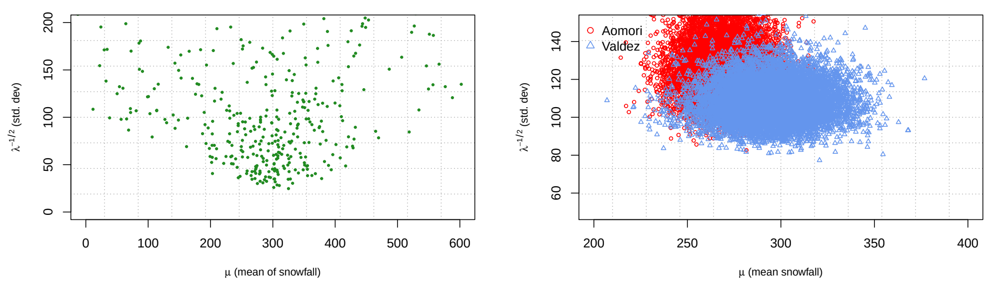




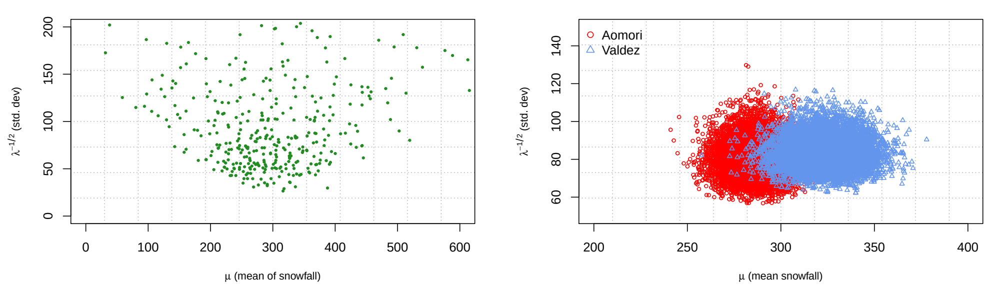
Setting $\mu_0 = 0$ change the location of of prior distribution. But it doesn't chage the posterior probability that the Valdez has a higher mean of snowfall is about 0.968.



Setting $\sigma \approx 10$ change the scale of prior distribution. But it doesn't chage the posterior probability that the Valdez has a higher mean of snowfall is about 0.968.



Setting $\nu = 4$ a high strenth of one observation to the information coming from the prior on μ_v and μ_a. It shows a higher stand diviation but it doesn't chage the posterior probability that the Valdez has a higher mean of snowfall is about 0.857.



Setting $\nu = 1/4$ a low strenth of one observation to the information coming from the prior on μ_v and μ_a. It shows a lower stand diviation but it doesn't chage the posterior probability that the Valdez has a higher mean of snowfall is about 0.988.