STAT 431 — Applied Bayesian Analysis — Course Notes

Outliers

Fall 2022

What if data suggest that a stochastic value in a model is unusual, relative to what is typical for the model?

It could be a data value y_i, or it could be a parameter θ_j in a hierarchical model.

In either case, it could be called an **outlier**.

Possible reasons: mistakes in the data, unused explanatory variables, mis-specified distributions

Typically, outliers in a quantitative variable are extreme (very large or small) compared to what the posterior would predict.

Models that use the normal distribution are especially susceptible to outliers because the normal density has **light (short) tails**:

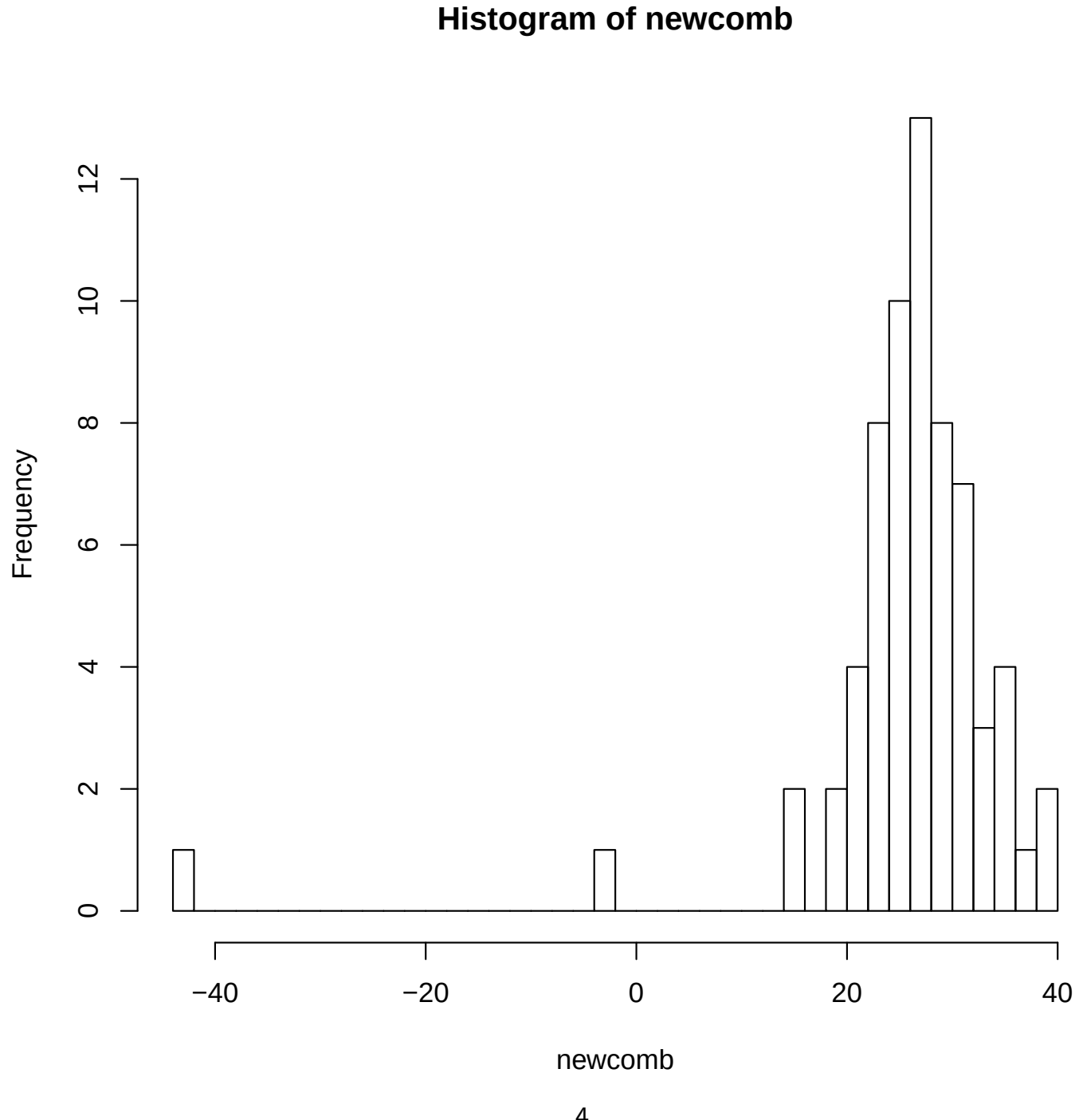
k	Prob. a $\mathrm{Normal}(\mu, \sigma^2)$ variable is $> k\sigma$ from μ
1	0.3173105
2	0.0455003
3	0.0026998
4	0.0000633
5	0.0000006

Newcomb Data

The `newcomb` data set comes from an early (1882) experiment to determine the speed of light. The 66 measurements are shifted and scaled times for light to travel the same known distance (in air).

Continuous measurements like these would typically be modeled with a normal distribution.

But the `newcomb` data have obvious outliers, relative to what would be expected under a normal distribution ...


Histogram of newcomb
Frequency
0
2
4
6
8
10
12
−40
−20
0
20
40
newcomb

Robustness

For models with standard distributions (e.g., normal), outliers can be highly influential — removing them disproportionately affects the inference.

We seek models that are more **robust**: less sensitive to extreme values of a few observations (or parameters).

A robust model can be used as an alternative to a standard model, either in sensitivity analysis, or for improved inference.

t-Distribution

Recall the (location-scale) t-distribution $\mathrm{t}_\nu(\mu, \sigma^2)$, with PDF

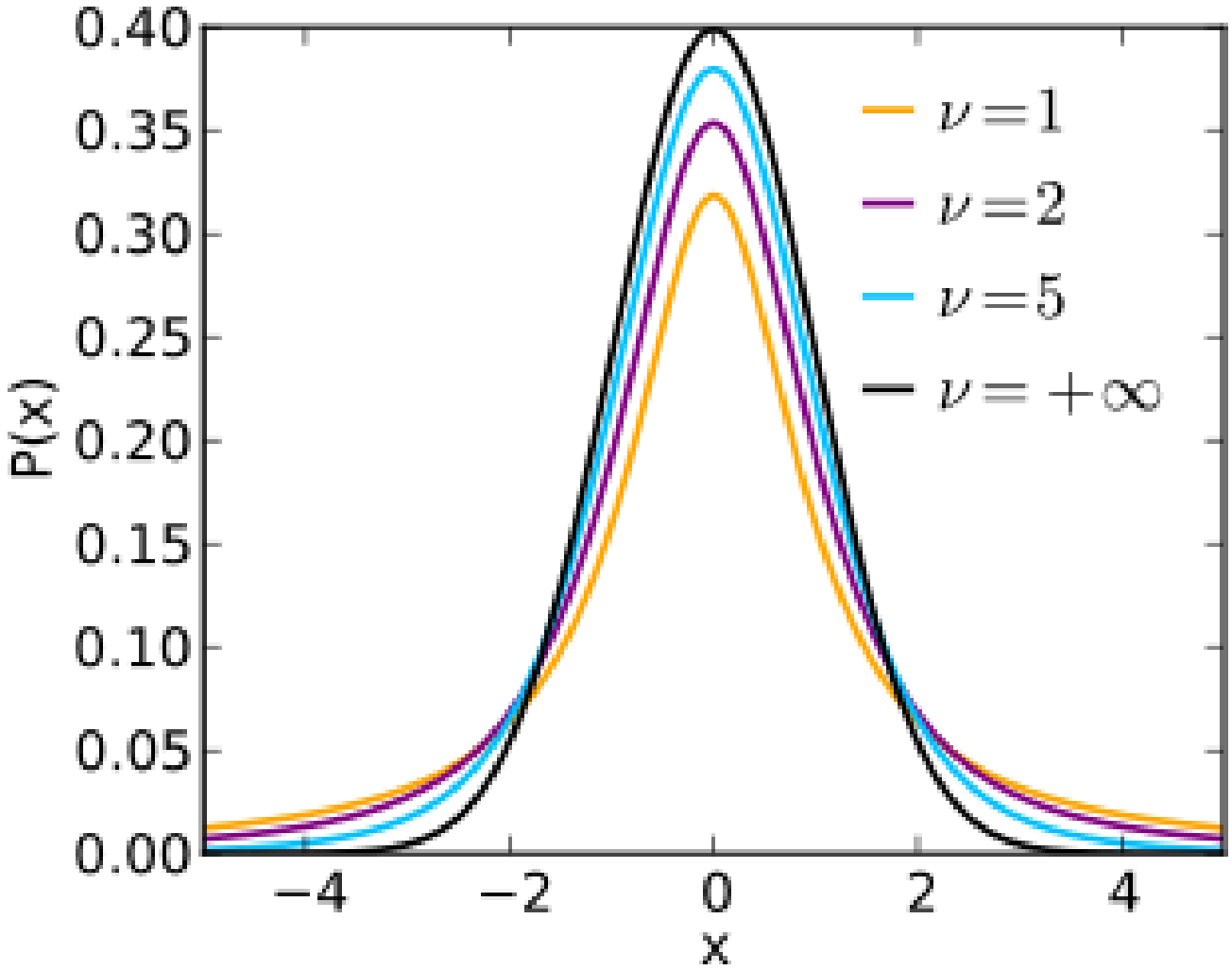
$$f(x) \quad \propto \quad \left(1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-(\nu+1)/2} \qquad \text{for all } x$$

The mean is μ when $\nu > 1$.
Otherwise, the mean is undefined, but μ is still the median.

The variance is

$$\frac{\nu}{\nu - 2}\,\sigma^2$$

if $\nu > 2$, and undefined otherwise.



From: Student's t-distribution. (2017, November 30). In *Wikipedia, The Free Encyclopedia*. Retrieved November 30, 2017, from `https://en.wikipedia.org/w/index.php?title=Student%27s_t-distribution&oldid=812931585`

For small ν, the t-distribution has **heavy (long) tails**, which are thicker than those of a normal.

For example, consider $\nu = 2$:

k	Prob. a $\mathrm{t}_2(\mu, \sigma^2)$ variable is $> k\sigma$ from μ
1	0.4226
2	0.1835
3	0.0955
4	0.0572
5	0.0377

For larger ν, the tails become lighter, and the distribution becomes more like a normal. In fact,

$$\mathrm{t}_\nu(\mu, \sigma^2) \quad \underset{\nu \to \infty}{\longrightarrow} \quad \mathrm{Normal}(\mu, \sigma^2)$$

So a t-distribution could have heavy or light tails, as controlled by ν.

Note: ν need not be an integer — it can be any positive value.

Heavy Tails and Outliers

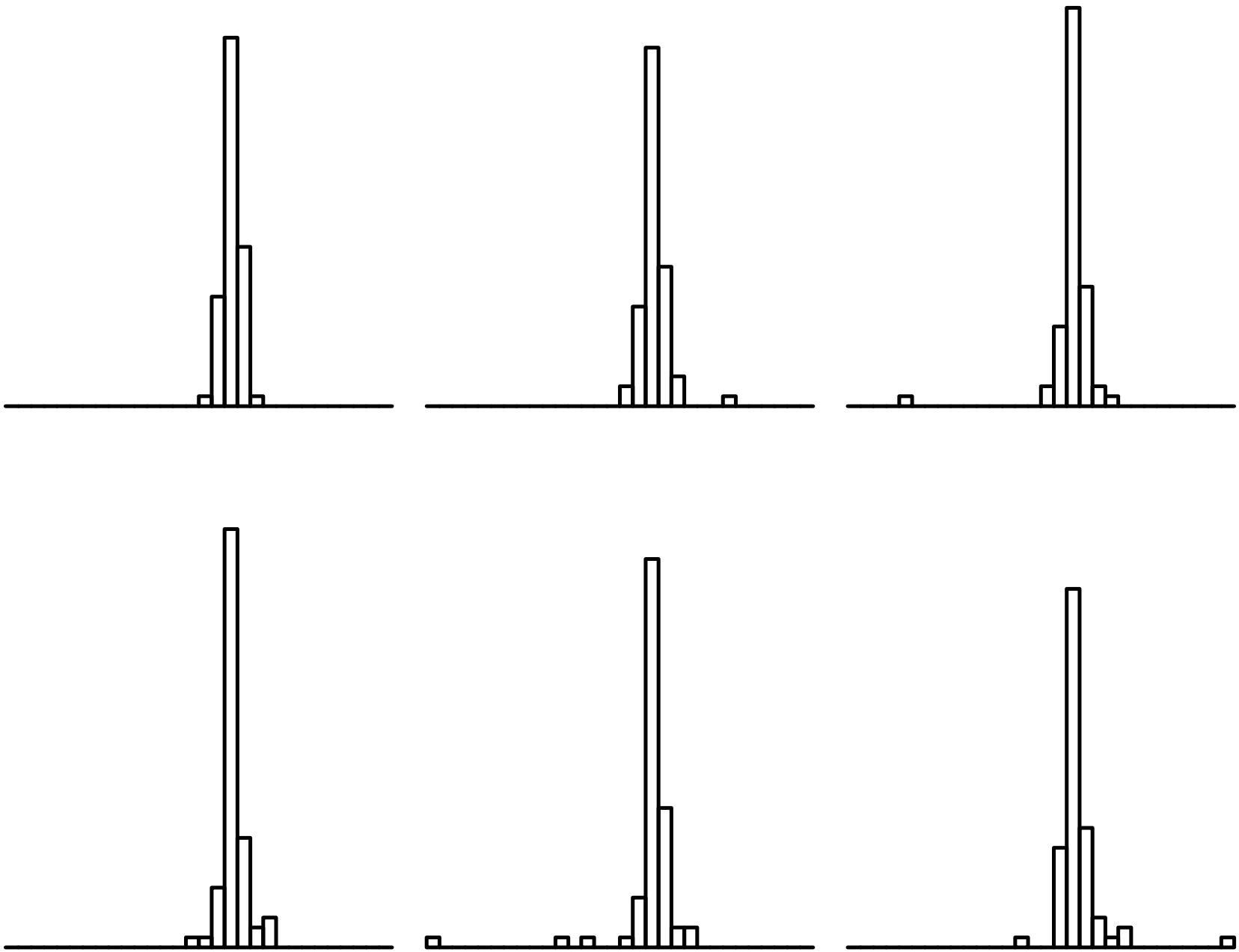
A continuous distribution with heavy tails tends to produce samples that have outliers, relative to a normal distribution.

For example, let's simulate 6 samples of size $n = 66$ from $\mathrm{t}_2(0, 1)$:

```
> n <- 66
```

```
> tsamp <- matrix(rt(6*n, 2), n, 6)
```

Now make a histogram of each column ...



Modeling with the t-Distribution

When the ideal data model would be a normal distribution, but the data have outliers, a (location-scale) t-distribution is a viable alternative.

Using the t requires a way to handle the degrees of freedom parameter $\nu > 0$. Smaller values give heavy tails, larger values give light tails.

A Bayesian model for a t-distributed sample:

$$
\begin{aligned}
Y_1, \ldots, Y_n \mid \mu, \sigma^2 &\sim iid\ \mathrm{t}_\nu(\mu, \sigma^2) \\
\mu &\sim 1\, d\mu \qquad\qquad -\infty < \mu < \infty \\
\sigma^2 &\sim \frac{1}{\sigma^2}\, d\sigma^2 \qquad\qquad \sigma^2 > 0
\end{aligned}
$$

ν is chosen arbitrarily, in this case.

Note: This uses the same "standard" improper joint prior as for a two-parameter normal sample.

Provided $\nu \geq 1$ and all $n > 1$ values y_i are distinct, the posterior is proper.

(Otherwise, be careful!)

Generally, we consider only $\nu > 1$, to retain interpretation of the mean (and to avoid possible issues with propriety and computation).

Random ν

To avoid specifying ν, we can give it a prior.

Because t converges to normal as $\nu \to \infty$, instead consider the reciprocal $1/\nu$:

$$0 \quad < \quad 1/\nu \quad < \quad 1$$

The lower bound represents the normal and the upper bound the (location-scale) Cauchy.

One textbook suggests

$$1/\nu \quad \sim \quad \mathrm{Uniform}(0, 1)$$

The resulting Bayesian model for a t-distributed sample:

$$
\begin{aligned}
Y_1, \ldots, Y_n \mid \mu, \sigma^2, \nu &\sim iid\ \mathrm{t}_\nu(\mu, \sigma^2) \\
\mu &\sim 1\, d\mu && -\infty < \mu < \infty \\
\sigma^2 &\sim \frac{1}{\sigma^2}\, d\sigma^2 && \sigma^2 > 0 \\
1/\nu &\sim \mathrm{Uniform}(0, 1)
\end{aligned}
$$

A Trick for Semi-Conjugacy

What if we want to use *proper* priors, instead?

The t-distribution has no obvious conjugate or semi-conjugate priors.

However, we can recover semi-conjugacy for μ and σ^2 by using a special representation of the t ...

Recall: If

$$\begin{aligned} X \mid W = w &\sim \text{Normal}(\mu_0, w/\kappa) \\ W &\sim \text{InvGamma}(\alpha, \beta) \end{aligned}$$

then

$$X \sim \text{t}_{2\alpha}\big(\mu_0,\, \beta/(\alpha\kappa)\big)$$

Now take

$$X = Y_i \qquad W = W_i \qquad \mu_0 = \mu$$

$$\kappa = 1 \qquad \alpha = \nu/2 \qquad \beta = \nu\sigma^2/2$$

to obtain ...

$$Y_i \mid \mu, \sigma^2, \nu \;\sim\; \mathrm{t}_\nu(\mu, \sigma^2)$$

if

$$\begin{aligned} Y_i \mid W_i = w_i, \mu &\sim \mathrm{Normal}(\mu, w_i) \\ W_i \mid \sigma^2, \nu &\sim \mathrm{InvGamma}(\nu/2,\, \nu\sigma^2/2) \end{aligned}$$

This represents the t-distribution of Y_i as a *scale mixture of normals*, by introducing a new latent variable W_i.

In this new hierarchical representation, μ is just a normal mean. It can be shown that

$$\mu \;\sim\; \mathrm{Normal}(\mu_0, \sigma_0^2)$$

is semi-conjugate.

What about σ^2?

The hierarchy involves σ^2 only as a factor in the second ("β") parameter of an inverse gamma distribution.

It can be shown that the *gamma* distribution is semi-conjugate for that parameter, and hence (by scaling) for σ^2:

$$\sigma^2 \sim \text{Gamma}(\alpha_0, \beta_0)$$

Unfortunately, there is no natural continuous semi-conjugate prior for ν.

Overall, the (partially) semi-conjugate hierarchy becomes

$$Y_i \mid W_i = w_i, \mu \;\sim\; indep\ \text{Normal}(\mu, w_i)$$

$$W_i \mid \sigma^2, \nu \;\sim\; iid\ \text{InvGamma}(\nu/2,\, \nu\sigma^2/2)$$

$$\left.\begin{aligned} \mu &\sim \text{Normal}(\mu_0, \sigma_0^2) \\ \sigma^2 &\sim \text{Gamma}(\alpha_0, \beta_0) \\ 1/\nu &\sim \text{Uniform}(0, 1) \end{aligned}\right\} \text{independent}$$

JAGS does not require us to use this hierarchy — the t-distribution can be specified directly — but it motivates a reasonable choice of proper prior.

Example: Newcomb Data

The data are in the `newcomb` data set of R package `MASS`:

$$Y_i \;=\; i\text{th shifted, scaled speed of light measurement}$$

$$i \;=\; 1, \ldots, 66$$

$$\bar{y} \;\approx\; 26.212 \qquad \text{median} \;=\; 27$$

Without the two outliers:

$$\bar{y} \;=\; 27.75 \qquad \text{median} \;=\; 27.5$$

A JAGS model for the case of random ν:

```
model {
  for(i in 1:length(y)) {
    y[i] ~ dt(mu, 1/sigmasq, 1/nuinv)
    yrep[i] ~ dt(mu, 1/sigmasq, 1/nuinv)
  }

  mu ~ dnorm(0, 0.00000001)
  sigmasq ~ dgamma(0.00001, 0.00001)
  nuinv ~ dunif(0, 1)
}
```

Note inclusion of a replicate data vector `yrep` for posterior predictive assessment.

R/JAGS Extra Example 1:

Robust t Location-Scale Analysis

Remarks:

- Can extend to linear regression: Replace normally-distributed observations with t-distributed observations.
- Can use the t-distribution in the prior portion of a hierarchy (e.g., to handle outliers in random effects).