Comparing Groups and Analyzing Experiments

School of Information Studies
Syracuse University

Learning Topics for This Week

Reinforcement of three perspectives on inference

Introduction to Analysis of Variance (ANOVA)

- Notes on experimental design
- Using variances to understand differences in means

Frequentist approach to ANOVA

- Between groups variance; within groups variance
- The F-distribution

Degrees of freedom

Bayesian approach to ANOVA

Bayes factors

Putting the evidence together

Reinforcing Three Perspectives

Three Perspectives on Inference

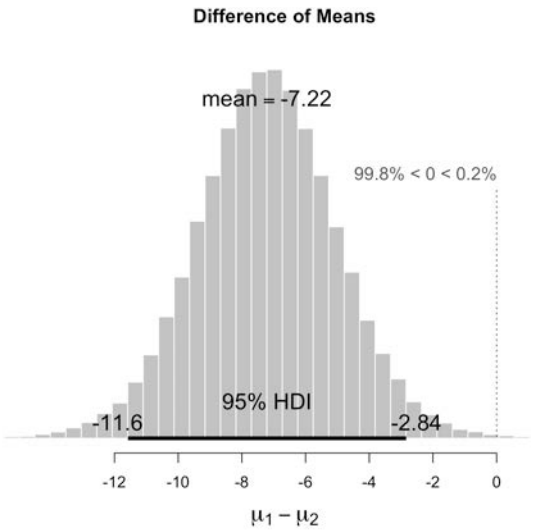
In all three cases we are trying to use sample data to obtain an estimate of the underlying population parameter; in some cases we want to know if the population parameter is, in effect, zero (e.g., the difference between two means)

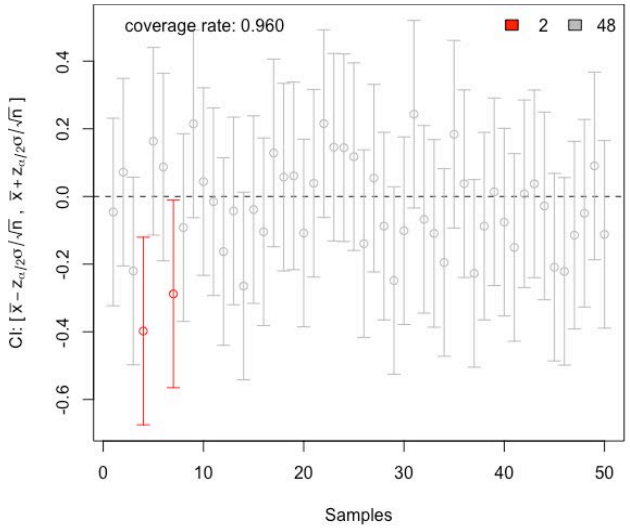
Bayesian: model the most likely position of the parameter by generating a posterior distribution

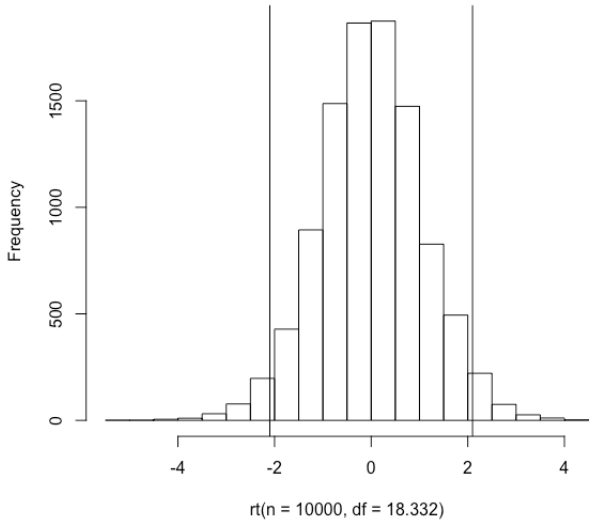
Confidence interval: build an interval around a point estimate whose width reflects the uncertainty of that estimate

Null Hypothesis Test: make a go/no-go decision by positioning a point estimate on a statistical model of the null hypothesis

Three Perspectives: Graphically







Introduction to Analysis of Variance

Notes on Experimental Design

Experiments provide the strongest evidence about causality

An experiment contains at least one "manipulation" or "treatment" that divides cases into at least two groups; in data this is represented as a categorical variable

The t-test (last week) is used to compare two groups, typically a control group and a treatment group

More complex experimental designs may have several treatment groups: for example, the effects of six different kinds of feed on the growth of chickens

Experiments in the technology arena are sometimes called A/B tests

Extending Comparison of Groups to More Than Two Groups

	t-Test	Analysis of Variance (ANOVA)
Purpose	Comparing two groups	Comparing two or more groups
Dependent/outcome variable	Any metric variable (also known as interval or ratio)	Any metric variable (also known as interval or ratio)
Independent variable	Unstated (technically membership in one or the other group)	Categorical (group designator)
Distribution for significance testing	The "t" distribution—looks like the normal distribution with somewhat heavy tails	The "F" distribution—asymmetric with a peak near one and a long positive tail
Bayesian approach	Create a posterior distribution of mean differences between groups	Create posterior distributions showing each group's differences from grand mean

Note: ANOVA can be used to analyze any categorical independent variable's effect on a metric dependent variable; the categorical variable may be naturally occurring, rather than a manipulation

Metric vs. Categorical

The "dependent" variable in ANOVA is always a "metric" variable—a measurement of something on a scale

- Examples of metric scales: temperature, height, weight, clicks per hour, ratings on a 10 point scale, SAT verbal score

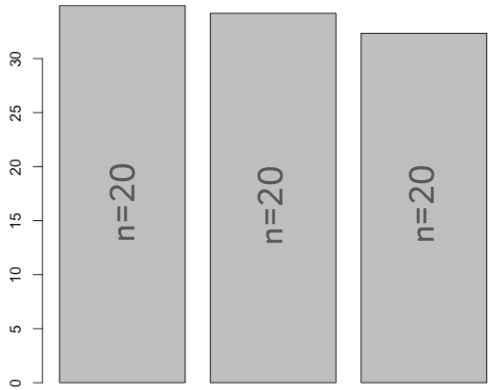
Every ANOVA also has an "independent" variable that is categorical—for purposes of analysis, all observations are divided into two or more mutually exclusive categories

- Examples of categorical variables: adult/child, Canada/Mexico/U.S., urban/suburban/rural, car/bus/truck

Categorical independent variables are sometimes referred to as "factors" and the various category options as "levels"

Using Variances to Understand Differences Among Means

Three groups randomly sampled from the same population



Variance among means: 1.78

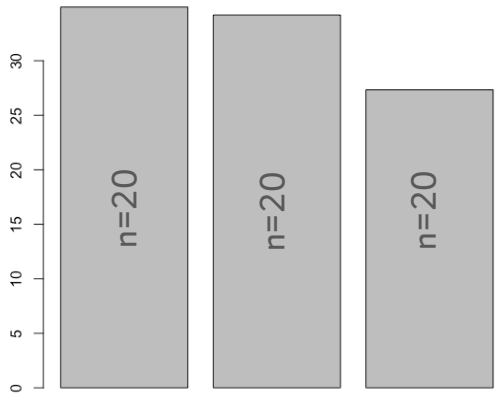
The bar chart on the left represents the means of three groups of n=20 measurements randomly sampled from a single (fictional) population of yearly rainfall measurements. The means are all close to 33.

The bar chart on the right is the same data, except that in the third group (rightmost), we subtracted 5 from every data point to drop the group mean down to about 28.

Think of that as a simulation of a drought in the third geographical region.

Note the difference in variances among the means.

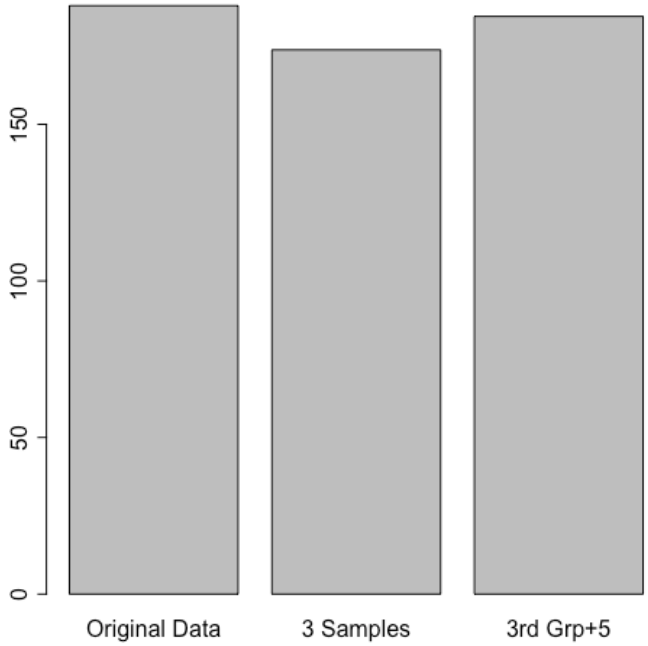
Subtract 5 from each member of right hand group



Variance among means: 17.5

Within Groups Variance

```
set.seed(1)
pgrp1 <- sample(precip,20, replace=TRUE)
pgrp2 <- sample(precip,20, replace=TRUE)
pgrp3 <- sample(precip,20, replace=TRUE)
v1 <- var(precip)
v2 <-var(c(pgrp1,pgrp2,pgrp3))
pgrp3 <- pgrp3 - 5
v3 <- var(c(pgrp1,pgrp2,pgrp3))
barplot(c(v1,v2,v3),names.arg=c("Original Data","3 Samples","3rd Grp+5"))
```



Frequentist Approach to ANOVA

ANOVA Null Hypothesis Test

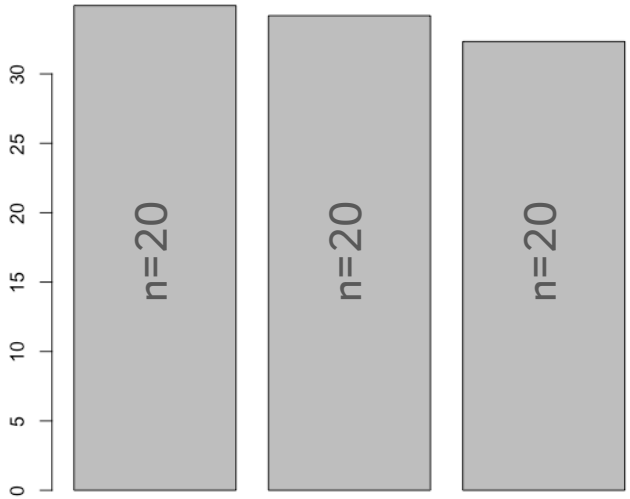
Compare three precipitation groups sampled from the same data

Minimal variance among means

ANOVA should show no meaningful differences among means

Null Hypothesis Significance test—should be “fail to reject” the null

Three groups randomly sampled from the same population



Variance among means: 1.78

The ANOVA Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
precipGrp	2	90	45.11	0.247	0.782
Residuals	57	10404	182.53		

Df: degrees of freedom—elements of a set that are free to vary once some statistics have been calculated; from a data set of 60 we lose 1 degree of freedom for calculating the grand mean; among the 3 group means only two can vary freely; this leaves 57 df within groups

Sum Sq: sum of squares—a raw initial calculation of variability; the first line is the "between groups" sum of squares discussed above; the second line is the "within groups" sum of squares

Mean Sq: mean squares, a.k.a. variance—the first line is the "between groups" variance as discussed above; the second line is the "within groups" variance

The ANOVA Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
precipGrp	2	90	45.11	0.247	0.782
Residuals	57	10404	182.53		

F value: the F-Ratio—quite literally a ratio of the mean squares from the first line (between groups) and the mean squares of the second line (within groups), i.e., 43.47 divided by 180.19

Pr(>F): the probability of a larger F-ratio—when examining a distribution of F-Ratios for the degrees of freedom appearing in this table, this is the probability of finding an F value at least this high (in this case at least 0.247); the F-distribution only has a positive tail, so to reject the null hypothesis, we must look for extreme values of F that appear in the tail of the distribution

The F-Ratio: A Ratio of Between Groups Variance vs. Within Groups Variance

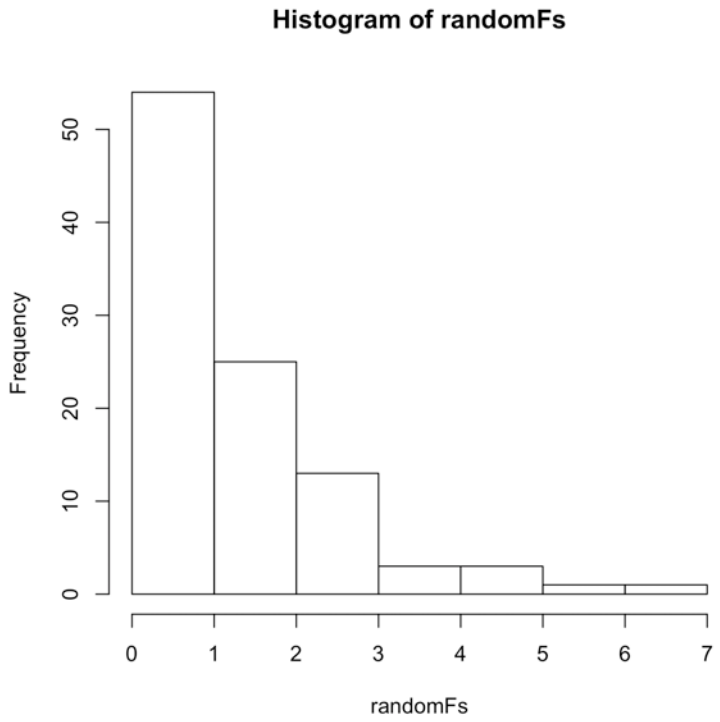
Between groups variance is the variation among the group means.

We obtain within groups variance by mixing together all of the data from the samples and obtaining the variance among these scores.

With a little adjusting to consider sample size, we can create a ratio of between groups variance to within groups variance.

Both are estimates of the population variance and if the groups were all sampled from the same population, the ratio should be close to 1.

F is a distribution of random values where the typical value is close to 1, although some values are larger than 1 because of sampling error.

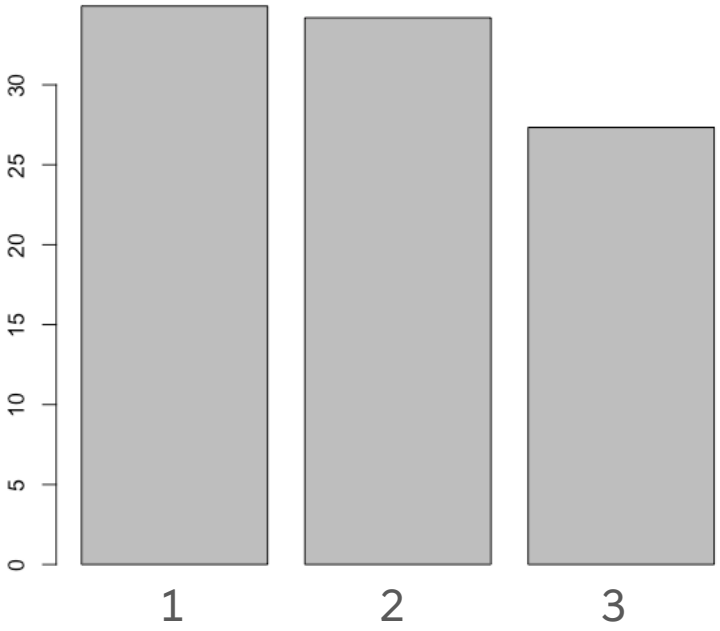


Post Hoc Testing

When a null hypothesis test of F is statistically significant, this result says nothing about which means are actually different from one another

In our three rainfall groups, if the F-test was significant with the means shown on the graph to the right, that would suggest that Group 3 is different from Group 1, Group 2, or both

We would need to conduct a "post hoc" test to see if Group 1 was different from Group 2



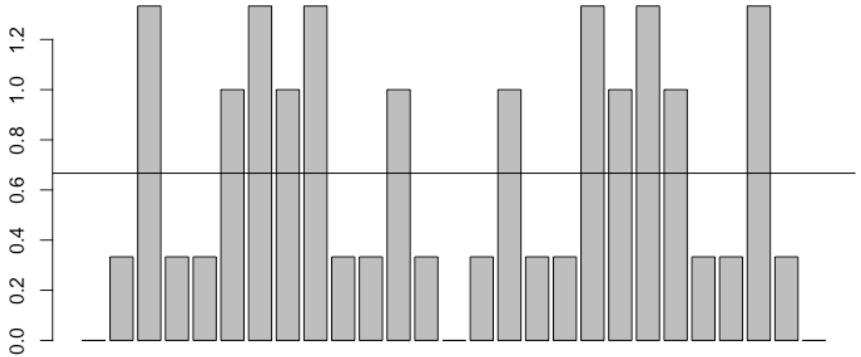
Introduction to Degrees of Freedom

Degrees of Freedom

A statistical concept having to do with the amount of information in a vector of numbers

When we calculate a statistic, such as the mean and use that in later calculations, we have borrowed part of our data

We reduce degrees of freedom to reflect this: e.g., df = (n - 1)

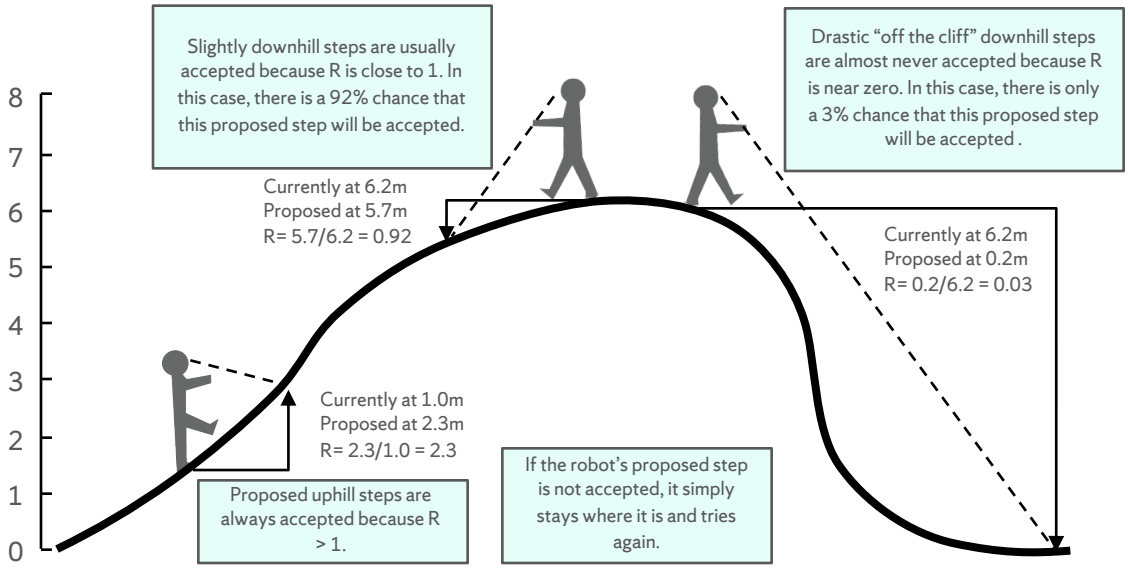


Variances of tiny samples, using (n - 1) in denominator; average across all 27 samples is identical to the true population variance

	[,1]	[,2]	[,3]
[1,]	1	1	1
[2,]	1	1	2
[3,]	1	1	3
[4,]	1	2	1
[5,]	1	2	2
[6,]	1	2	3
[7,]	1	3	1
[8,]	1	3	2
[9,]	1	3	3
[10,]	2	1	1
[11,]	2	1	2
[12,]	2	1	3
[13,]	2	2	1
[14,]	2	2	2
[15,]	2	2	3
[16,]	2	3	1
[17,]	2	3	2
[18,]	2	3	3
[19,]	3	1	1
[20,]	3	1	2
[21,]	3	1	3
[22,]	3	2	1
[23,]	3	2	2
[24,]	3	2	3
[25,]	3	3	1
[26,]	3	3	2
[27,]	3	3	3

Bayesian Approach to ANOVA

Reminder of MCMC: Robot Analogy



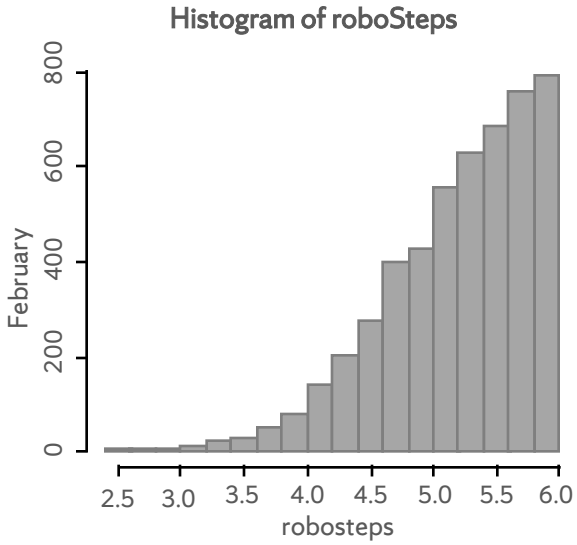


Image credit: Paul O. Lewis, http://marple.eeb.uconn.edu/mcmcrobot/?page_id=24

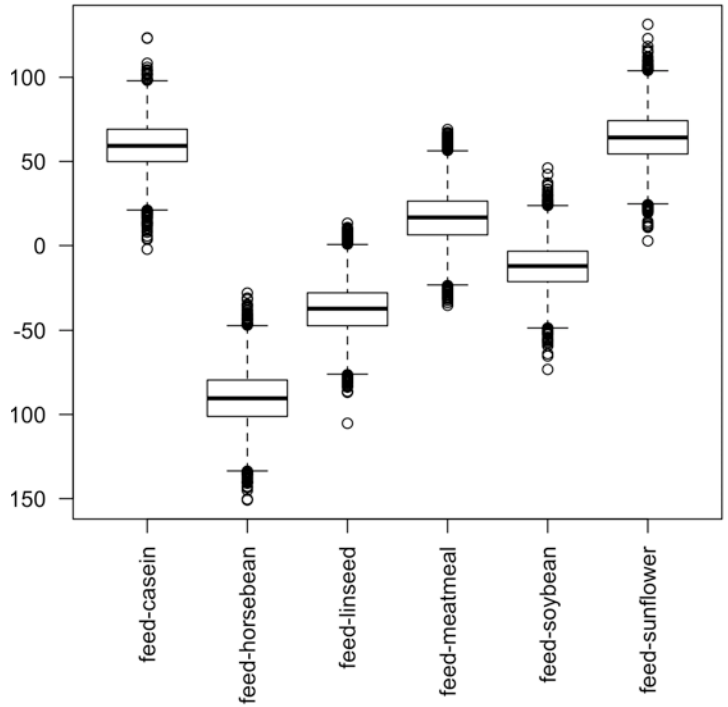
The Bayesian Approach to ANOVA

These results come from the chickwts dataset: 71 measurements of the weights of six-week old chicks fed on different diets.

The boxplot at left shows posterior distributions for deviations of each group from the grand mean (which is about 259 grams). The grand mean is represented on this figure as 0 on the y-axis.

The detailed numeric output from anovaBF() provides the specific boundaries of the 95% highest density intervals (HDIs) around each group deviation (see next slide).

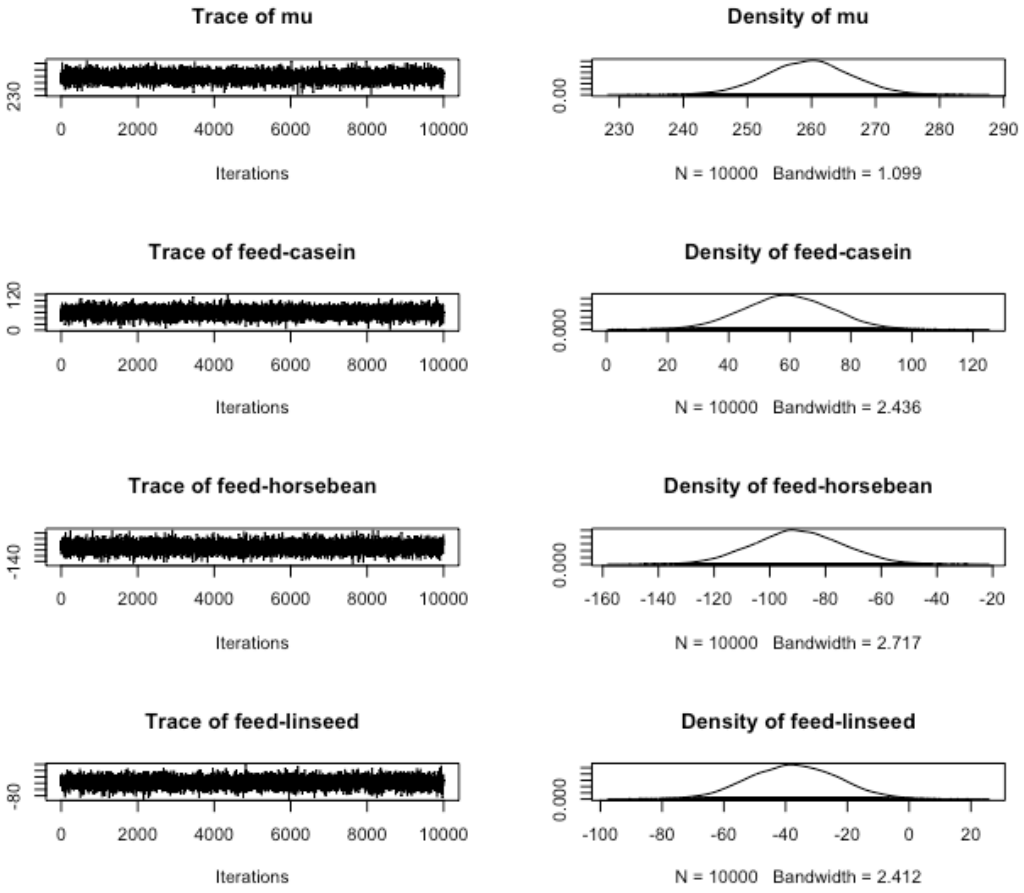
As a rule of thumb, pay attention to the outer whiskers of each box. Where there is no overlap, group means are substantially different.



Interpreting the HDI Table

	2.5%	25%	50%	75%	97.5%
mu	246.1102	254.7790	259.236	263.884	272.637
feed-sunflower	35.5407	54.4183	64.198	74.220	93.330
feed-casein	30.4805	49.8965	59.270	69.119	87.816
feed-meatmeal	-12.6600	6.3583	16.712	26.414	45.732
feed-soybean	-38.4556	-21.4666	-12.178	-3.364	14.310
feed-linseed	-65.9157	-47.2791	-37.497	-28.080	-9.537
feed-horsebean	-122.0330	-101.1611	-90.387	-79.524	-58.058

Trace Plots From Bayesian ANOVA



Bayes Factors

Bayes Factors: Odds in Favor of a Hypothesis

```
chicksBayesOut <- anovaBF(weight ~ feed, data=chickwts)

# Yields a Bayes Factor of 14067867:1 in favor of an effect for feed type—pretty strong!

# Technically, odds ratio represents comparison between alternative hypothesis and null

# Kass and Raftery (1995) rules of thumb:

# any odds ratio weaker than 3:1 is not worth mentioning

# odds ratios from 3:1 up to 20:1 are positive evidence for the favored hypothesis

# odds ratios from 20:1 up to 150:1 are strong evidence

# odds ratios of more than 150:1 are very strong evidences for the favored hypothesis
```

Putting the Evidence Together

Putting All of the Evidence Together

The F Test (Frequentist)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
feed	5	231129	46226	15.37	5.94e-10 ***
Residuals	65	195556	3009		

The Bayes Factor

[1] feed : 14067867 ±0%

Against denominator:

Intercept only

The HDIs for Group Mean Deviations (from MCMC; showing just three groups)

	2.5%	25%	50%	75%	97.5%
feed-sunflower	35.54	54.4183	64.20	74.22	93.330
feed-casein	30.48	49.8965	59.27	69.12	87.816
feed-horsebean	-122.03	-101.16	-90.387	-79.52	-58.058

Example: Written Interpretation of Chick Weight Results

An experiment was conducted by feeding one of six different diets to n=71 chicks. Results of a conventional ANOVA showed a significant difference among weights of chicks measured at six-weeks old ($F(5,65)=15.37$, $p<.001$). A Bayesian test confirmed this result with a Bayes factor of 14067867:1 in favor of a mean differences model as compared to an intercept-only model.

A Bayesian analysis of the group means showed that both sunflower and casein (milk protein) provided superior diets with an estimated 64.2 gram and 57.3 respective increases over the overall mean weight of 259.2 grams. Highest density intervals of the posterior distributions for these means overlapped slightly with meat meal, but had no overlap (and therefore were superior to) soybean, linseed, and horsebean diets.