# L5B: Hypothesis Testing

## Overview

Hypothesis testing is a formal procedure for comparing competing theories about natural phenomena. It can be viewed as a key component of the scientific method and, in general, a means of advancing knowledge and understanding.

The simplest scenario has two competing hypotheses, one labelled the Null Hypothesis and denoted $H_0$ and the other labelled the Alternative Hypothesis and denoted $H_1$. In our statistical framework these hypotheses are typically statements about the possible values of a parameter (or parameters):

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1$$

The sets defined by the hypotheses are mutually exclusive, $\Theta_0 \cap \Theta_1 = \varnothing$, and (usually) exhaustive, i.e., their union includes the entire parameter space, $\Theta_0 \cup \Theta_1 = \Theta$.

**Example.** The fraction of a specific variety of potatoes infected by a virus is approximately 0.15. A virus resistant variety of potatoes will be planted this year and the hope is that that fraction infected will be less than 0.15. Letting $\theta$ denote the probability that plant is infected, the two competing hypotheses are:

$$H_0 : \theta \geq 0.15$$
$$H_1 : \theta < 0.15$$

**Example.** In multiple regression one is interested in knowing if one or more covariates have a linear relationship with a response variable. For example, is this model, $E[Y]=\beta_0 + \beta_1 x_1 + \beta_2 x_2$, correct? Two hypotheses might be:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

**Comments**

- A Null Hypothesis that includes only a single value for $\theta$ it is called a Point Null Hypothesis (or Simple Hypothesis). There are often practical problems with such hypotheses; e.g., $\beta_1 = 0.001$ would be contrary to $H_0$.
- Competing hypotheses can (sometimes) be viewed as competing models about phenomena. The above hypotheses could be written as:

  $$H_0 \equiv M_0 \text{ is the correct model} : E[Y] = \beta_0 + \beta_2 x_2$$
  $$H_1 \equiv M_1 \text{ is the correct model} : E[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

  And one can easily imagine a larger set of hypotheses or alternative models:

  $$M_1 : E[Y] = \beta_0$$
  $$M_2 : E[Y] = \beta_0 + \beta_1 x_1$$
  $$M_3 : E[Y] = \beta_0 + \beta_2 x_2$$
  $$M_4 : E[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

# 1 Classical Hypothesis Testing

The classical or frequentist approach to testing two hypotheses is:

- Assume that one hypothesis is true, $H_0$.
- Calculate a test statistic based on the observed sample data, $T(\mathbf{y}_{obs})$ (that should be informative about $H_0$ and $H_1$).
- *Conditional* on $H_0$ being true, calculate the probability of observing sample data that would yield test statistics as extreme or more extreme than $T(\mathbf{y}_{obs})$.

  That probability is called the *p-value* and is formally defined:

$$p - value = \Pr(T(\mathbf{y}) \text{ more extreme than } T(\mathbf{y}_{obs})|\theta, H_0) \tag{1}$$

  where "extremeness" is in the direction of the alternative hypothesis.
- If that probability is
  - "sufficiently small", "Reject $H_0$" and "Accept $H_1$".
  - "relatively large", "Do not reject $H_0$." (*But Do Not Say* "Accept $H_0$.")

**Example A.** The sampling model is Normal($\theta$, $\sigma^2$), where $\theta$ is unknown but $\sigma^2$ is known and equals 2 (admittedly seldom realistic). The null hypothesis is that $\theta$ is less than or equal to 3 while the alternative hypothesis is that $\theta$ is greater than 3:

$$H_0 : \theta \leq 3$$
$$H_1 : \theta > 3$$

A random sample of $n$=10 is taken and the sample average is $\bar{y} = 4$. The test statistic:

$$T(\mathbf{y}) = \frac{\bar{y} - \theta_0}{\sqrt{\sigma^2/n}}$$

where $\theta_0$ is a value in the set $\theta \leq 3$. Note that conditional on $H_0$, $T(\mathbf{y})$ is Normal(0,1)[1]. Given that there are infinite number of values in the set $\Theta_0$, the convention is to select the value of $\theta \in \Theta_0$ that would yield the largest p-value, in this case $\theta_0$=3, and then the p-value is

$$\Pr\left(T(\mathbf{y}) \geq T(observed)\right) = \Pr\left(T(\mathbf{y}) \geq \frac{4-3}{\sqrt{2/10}} = 2.236\right) = 1 - \Phi(2.236) = 0.013$$

where $\Phi(z)$ is the cumulative distribution function for a standard normal random variable. Note that extremeness here is in the direction of $H_1$, namely, towards values of $\theta > 3$. Such a p-value of 0.013 would be considered by many to be "sufficiently small", or *statistically significant*, and $H_0$ would be rejected.

**Example B.** Two linear models for an expected outcome are proposed, where one model is nested inside the other model:

$$M1 : E[Y] = \beta_0 + \beta_1 x$$
$$M2 : E[Y] = \beta_0 + \beta_1 x + \beta_2 x^2$$

[1]The test statistic $T(\mathbf{y})$ for this setting is sometimes written $z$ and is called the $z$-statistic.

Equivalently,

$$H_0 : \beta_2 = 0$$
$$H_1 : \beta_2 \neq 0$$

Assuming normality of $Y$, the common test statistic is the $t$-statistic, $t = \frac{\hat{\beta}_2 - 0}{std.error(\hat{\beta}_2)}$. And extremeness in this case would be values of $t$ that are relatively far from 0, $t << 0$ or $t >> 0$.

**Problems with classical hypothesis testing.**

1. $H_0$ and $H_a$ must be structured that "extremeness" in the direction of $H_a$ is definable in order to calculate the p-value. If one is comparing models that are not nested, "extremeness" is not readily definable. For example, exponential "growth" versus linear "growth" models:

   $$M1 : E[Y] = \beta_0 \exp(\beta_1 t)$$
   $$M2 : E[Y] = \beta_0 + \beta_1 t$$

   If $H_0$ is that $M1$ is true, and $H_1$ is that $M2$ is true, then assuming that $H_0$ is true, what is a measure of extremeness in the directon of $H_1$?

2. The evidence is only *against* $H_0$ as the p-value is calculated *assuming* that $H_0$ is true.

   - A small p-value indicates that the data are not what would be expected if $H_0$ is true.
   - A large p-value, however, does not mean that $H_0$ is true, that the model implied by $H_0$ is true, as the calculation is made *assuming* that $H_0$ is true—-so there is no weight of evidence *for* $H_0$.
   - This is the reason that the frequentist conclusion given a large p-value is to say "fail to reject" $H_0$, and *Not* to say "accept" $H_0$. You can't accept something that you assumed was true in the first place.

3. The p-value itself, e.g., 0.01, does not provide "weight of evidence" for the $H_0$. The p-value is a long-run relative frequency measure: if $H_0$ was true, only 1% of the time would the observed results or *more extreme* results. The p-value is not the probability that $H_0$ is true.

4. Calculation of P-values involves including values that were not even observed. This violates the Likelihood Principle[2].

   **Example C.** (This example was discussed previously in Lecture Notes 1.) The sampling model for the data is Poisson($\theta$) and there are two hypotheses about $\theta$:

   $$H_0 : \theta = 1 \qquad H_1 : \theta = 2$$

   A sample size $n$=1 is drawn and yields the value $y$=2. The standard frequentist approach is to calculate the p-value: the probability of the observed value and any values in a direction away from $H_0$ in the direction of $H_1$. In this case the p-value is $\Pr(Y \geq 2|H_0)$ = 1- $\Pr(Y = 0 \cup Y = 1|\theta = 1) = 0.264$[3]. Thus, one would not reject $H_0$.

   This procedure is violating the Likelihood Principle, however, in that inference is being based on more than the likelihood of the data: the probability of events that *did not occur*, such as $Y$=3 or $Y$=4, is being used as the basis for inference.

[2]Reminder from Lecture 1 Notes: The Likelihood Principle says that given a sample of data, $\mathbf{y}$, any two sampling models for $\mathbf{y}$, say $p_1(\mathbf{y}|\theta)$ and $p_2(\mathbf{y}|\theta)$, that have the same likelihood value yield the same inference for $\theta$. The main point is that inference for $\theta$ depends on the observed $\mathbf{y}$ alone, not on unobserved values of $\mathbf{y}$.

[3]In R: `1-ppois(q=1,lambda=1)`=0.2642411.

# 2 Bayesian Hypothesis Testing

Suppose that there are two hypotheses about a parameter $\theta$:

$$H_0 : \theta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \theta \in \Theta_1$$

where $\Theta_0 \cup \Theta_1$ is the entire parameter space and $\Theta_0 \cap \Theta_1 = \emptyset$.

The Bayesian approach is to specify prior probabilities on each hypothesis:

$$p_0 = \Pr(H_0 \text{ is true}) = \Pr(\theta \in \Theta_0)$$
$$p_1 = \Pr(H_1 \text{ is true}) = \Pr(\theta \in \Theta_1)$$

, and where $p_0 + p_1 = 1$.

Then data, $\mathbf{y}$, are collected and the posterior probabilities for each hypothesis are calculated:

$$\Pr(H_0|\mathbf{y}) = \Pr(\theta \in \Theta_0|\mathbf{y})$$

And $\Pr(H_1|\mathbf{y})$ = 1- $\Pr(H_0|\mathbf{y})$.

The complexity of the calculation of the posterior probability is affected by the nature of the hypotheses, i.e., simple or composite.

## 2.1 Simple

Simple hypotheses have single parameter values:

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta = \theta_1$$

Then

$$\Pr(H_0|\mathbf{y}) = \Pr(\theta = \theta_0|\mathbf{y}) = \frac{f(\mathbf{y}|\theta_0)p_0}{m(\mathbf{y})} = \frac{f(\mathbf{y}|\theta_0)p_0}{f(\mathbf{y}|\theta_0)p_0 + f(\mathbf{y}|\theta_1)p_1}$$
$$\Pr(H_1|\mathbf{y}) = \Pr(\theta = \theta_1|\mathbf{y}) = \frac{f(\mathbf{y}|\theta_1)p_1}{m(\mathbf{y})} = \frac{f(\mathbf{y}|\theta_1)p_1}{f(\mathbf{y}|\theta_0)p_0 + f(\mathbf{y}|\theta_1)p_1}$$

Given that $\Pr(H_0|\mathbf{y}) + \Pr(H_1|\mathbf{y}) = 1$, $\Pr(H_1|\mathbf{y})$ is simply $1 - \Pr(H_0|\mathbf{y})$.

Note that to calculate posterior odds the normalizing constant $m(\mathbf{y})$ need not be calculated:

$$\frac{\Pr(H_0|\mathbf{y})}{\Pr(H_1|\mathbf{y})} = \frac{f(\mathbf{y}|\theta_0)p_0}{f(\mathbf{y}|\theta_1)p_1}$$

## 2.2 Composite

Composite hypotheses include sets of parameter values:

$$H_0 : \theta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \theta \in \Theta_1$$

Letting $\pi(\theta)$ be the prior probability over the entire parameter space, the prior probability for Hypothesis $i$ is

$$p_i = \int_{\theta \in \Theta_i} \pi(\theta) \mathrm{d}\theta$$

Thus the prior distribution for the parameter $\theta$, $\pi(\theta)$, is *inducing* the prior for the hypothesis, $p_i$[4].

Now

$$\Pr(H_i|\mathbf{y}) = \frac{p(\mathbf{y}, H_i)}{m(\mathbf{y})} = \frac{p_i\ p(\mathbf{y}|H_i)}{m(\mathbf{y})} = \frac{p_i \int p(\mathbf{y}, \theta|H_i)\ \mathrm{d}\theta}{m(\mathbf{y})} = \frac{p_i \int f(\mathbf{y}|\theta)\pi(\theta|H_i)\ \mathrm{d}\theta}{m(\mathbf{y})}$$
$$= \frac{p_i \int_{\theta\in\Theta_i} f(\mathbf{y}|\theta)\frac{\pi(\theta)}{p_i}\ \mathrm{d}\theta}{m(\mathbf{y})} = \frac{\int_{\theta\in\Theta_i} f(\mathbf{y}|\theta)\pi(\theta)\mathrm{d}\theta}{m(\mathbf{y})} = \int_{\theta\in\Theta_i} p(\theta|\mathbf{y})\ \mathrm{d}\theta = \Pr(\theta\in\Theta_i|\mathbf{y})$$

The key step in the above is the equality of integrating $\pi(\theta|H_i)$ over the entire parameter space $\Theta$ and integrating $\pi(\theta)/p_i$ over the reduced parameter space $\Theta_i$.

Stated most plainly, however, the posterior probability of $H_i$ is simply the integral of the posterior for $\theta$ over $\Theta_i$.

The posterior odds of $H_0$ against $H_1$ can be written:

$$\frac{\Pr(H_0|y)}{\Pr(H_1|y)} = \frac{\int_{\theta\in\Theta_0} f(y|\theta)\pi(\theta)\mathrm{d}\theta}{\int_{\theta\in\Theta_1} f(y|\theta)\pi(\theta)\mathrm{d}\theta} = \frac{\Pr(\theta\in\Theta_0|y)}{\Pr(\theta\in\Theta_1|y)} = \frac{\Pr(\theta\in\Theta_0|y)}{1-\Pr(\theta\in\Theta_0|y)}$$

### 2.3 Remarks

- **Multiple Hypotheses.** Multiple hypotheses can be handled similarly. The different hypotheses could correspond to different sets of models: $M_1$, ..., $M_K$:

  $$H_i : \text{The correct model is model } M_i$$

  One would assign priors to each hypothesis, $p(H_i)$, where $\sum_{i=1}^K p(H_i) = 1$. Then the posterior probability for model $i$:

  $$\Pr(H_i|y) = \frac{\Pr(H_i, y)}{\Pr(y)} = \frac{\Pr(H_i, y)}{\sum_{j=1}^K \Pr(H_j, y)}$$

  where the form of $\Pr(H_i, y)$ would depend upon whether $H_i$ was simple or composite.

- **Computational difficulties.** For composite hypotheses, the integration needed to calculate $\Pr(\theta \in \Theta|y)$ may not be analytically tractable.

## 3 Bayes Factors

An alternative to calculating posterior probabilities for the hypotheses is Bayes factors. A Bayes factor is the ratio of posterior odds to prior odds. The *prior odds* for $H_0$ against $H_1$ is the ratio $p_0/p_1$. E.g., if $p_0$=0.6 and $p_1$=0.4, then $0.6/0.4 = 1.5$ are the prior odds. The *posterior odds* for $H_0$ against $H_1$ is the ratio $\Pr(H_0|\mathbf{y})/\Pr(H_1|\mathbf{y})$. The Bayes Factor for $H_0$ against $H_1$, which is written $BF_{01}$, is

$$BF_{01} = \frac{\Pr(H_0|y)/\Pr(H_1|y)}{p_0/p_1} = \frac{\Pr(\theta\in\Theta_0|y)/\Pr(\theta\in\Theta_1|y)}{p_0/p_1} \tag{2}$$

Rules of thumb for interpreting Bayes Factors are given by Kass and Raftery (Journal of the American Statistical Association Volume 90, 1995 - Issue 430 ):

[4]Note: one can specify a prior for the hypothesis independent of the prior for $\theta$; e.g., simply state that $p_0$=0.3 regardless of the $\pi(\theta)$.

| $BF_{01}$ | Interpretation |
|---|---|
| $< 3$ | No evidence for $H_0$ over $H_1$ |
| $> 3$ | Positive evidence for $H_0$ |
| $> 20$ | Strong evidence for $H_0$ |
| $> 150$ | Very strong evidence for $H_0$ |

Note: $BF_{10} = 1/BF_{01}$. And

- $BF_{01} < \frac{1}{3} \Rightarrow BF_{10} > 3 \Rightarrow$ positive evidence for $H_1$
- $BF_{01} < \frac{1}{20} \Rightarrow BF_{10} > 20 \Rightarrow$ strong evidence for $H_1$

### 3.1 Simple vs Simple

$H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$.

$$BF_{01} = \frac{\Pr(H_0|y)/\Pr(H_1|y)}{p_0/p_1} = \frac{f(y|\theta_0)p_0/f(y|\theta_1)p_1}{p_0/p_1} = \frac{f(y|\theta_0)}{f(y|\theta_1)} \tag{3}$$

Thus the Bayes Factor is simply the ratio of the likelihoods, and the priors for the hypotheses are irrelevant.

**Example C (continued).** The sampling distribution for the data is Poisson($\theta$) and $H_0 : \theta = 1$ and $H_1 : \theta = 2$. The prior for $H_0$ is $p_0$=0.8, thus $p_1$=0.2. A single observation, $n = 1$, is observed with $y = 2$. Then $BF_{01} = e^{-1}1^2/e^{-2}2^2 = 0.6796$, and $BF_{10} = 1.4715$. Thus there is no evidence for $H_0$ over $H_1$, or for $H_1$ over $H_0$.

### 3.2 Composite vs Composite

$H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$; $\Theta_0 \cup \Theta_1 = \Theta$.

$$BF_{01} = \frac{\Pr(H_0|y)/\Pr(H_1|y)}{p_0/p_1} = \frac{\left[\int_{\theta\in\Theta_0} f(y|\theta)\pi(\theta)\mathrm{d}\theta\right] / \left[\int_{\theta\in\Theta_1} f(y|\theta)\pi(\theta)\mathrm{d}\theta\right]}{p_0/p_1} = \frac{\Pr(\theta \in \Theta_0|y)/\Pr(\theta \in \Theta_1|y)}{p_0/p_1} \tag{4}$$

### 3.3 Simple vs Composite

$H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$.

$$BF_{01} = \frac{\Pr(H_0|y)/\Pr(H_1|y)}{p_0/p_1} = \frac{[f(y|\theta_0)p_0]/\left[p_1 \int_{-\infty}^{\infty} f(y|\theta)\pi(\theta)\mathrm{d}\theta\right]}{p_0/p_1} = \frac{f(y|\theta_0)}{\int_{-\infty}^{\infty} f(y|\theta)\pi(\theta)\mathrm{d}\theta} = \frac{f(y|\theta_0)}{m(y)} \tag{5}$$

Notes:

- As for the simple versus simple case, the prior probabilities for the hypotheses are cancelling out.

# 4 Example D: Simple Null and Simple Alternative

The number of hairs per square inch of mohair fabric used by a teddy bear manufacturer is assumed to have a Poisson($\theta$) distribution (King and Ross, 2017). The manufacturer wants to test the hypotheses:

$$H_0 : \theta = 100; \quad H_1 : \theta = 110$$

To test these hypotheses an independent random sample of $n$ pieces of fabric is drawn and the number of hairs per square inch, $\mathbf{y} = y_1, \ldots, y_n$, is recorded.

1. The Bayes Factor, $BF_{01}$:

$$\begin{aligned} BF_{01} &= \frac{\Pr(H_0|\mathbf{y})/\Pr(H_1|\mathbf{y})}{\Pr(H_0)/Pr(H_1)} = \frac{\Pr(\mathbf{y}|H_0)\Pr(H_0)/\Pr(\mathbf{y}|H_0)\Pr(H_0)}{\Pr(H_0)/Pr(H_1)} \\ &= \frac{\Pr(\mathbf{y}|H_0)}{\Pr(\mathbf{y}|H_1)} = \frac{\exp(-100*n)100^{n\bar{y}}}{\exp(-110*n)110^{n\bar{y}}} = \exp(10n)\left(\frac{100}{110}\right)^{n\bar{y}} \end{aligned}$$

2. Given $n$=10 and $\bar{y} = 102.7$:

$$BF_{01} = \exp(10*10)\left(\frac{100}{110}\right)^{10*102.7} = 8.301576$$

   which is between 3 and 20, thus "positive evidence" for $H_0$.

3. Assume the priors for $H_0$ and $H_1$ were $p_0$=$p_1$=0.5. The posterior probabilities for each hypothesis can be calculated directly from the Bayes Factor as follows:

$$\Pr(H_0|\mathbf{y}) = \frac{BF_{01}}{1+BF_{01}} = \frac{8.301576}{1+8.301576} = 0.8924913$$

   **Exercise:** show why the above formula works.

**Exercise:** Given $y_i$, $i$=1,...,10 are independent Normal($\mu$,1) random variables, where observe data:

```
3.4, 2.9, 3.0, 3.5, 3.3, 3.7, 2.7, 3.9, 2.7, 2.9
```

Test the simple hypothesis: $H_0 : \mu = 3$ vs $H_1 : \mu = 3.5$. Show that the Bayes Factor $BF_{01} = 1.28$.

# 5 Example E: Composite Null and Composite Alternative

A food manufacturer is considering releasing a new flavour of hummus, but before doing so wants to carry out an experiment with volunteers to see whether this new flavour is liked better than a competitor's version (based on example from Carlin and Louis, 2009). They would like to be "pretty sure" that the new flavour is preferred by at least 60% of hummus consumers. Letting $\theta$ be the probability that the new flavour is preferred, there are two hypotheses:

$$H_0 : \theta \geq 0.6 \quad vs \quad H_1 : \theta < 0.6$$

If there is strong evidence for $H_0$, they will release the new flavour. The manufacturer would prefer to be cautious and selects a Beta prior for $\theta$ that has an expected value of 0.5 and a coefficient of variation of 0.3 (thus a standard deviation of 0.5*0.3=0.15). That translates into Beta(5.056, 5.056).

The *induced* prior probability for $H_0$ is then[5]:

$$p_0 = \int_{0.6}^{1} \frac{1}{Be(5.056, 5.056)} \theta^{4.06}(1-\theta)^{4.06} \mathrm{d}\theta = \int_{0.6}^{1} \frac{\Gamma(10.13)}{\Gamma(5.065)\Gamma(5.065)} \theta^{4.06}(1-\theta)^{4.06} \mathrm{d}\theta = 0.265$$

Thus $p_1 = 0.735$.

To test these hypotheses, a taste preference study is carried with $n$=16 volunteers. How would you recommend that such a study be carried out?

Assume that the probability of preferring the new flavour is the same for all volunteers and the responses are independent. Then, letting $y$ be the number preferring the new flavour, $y \sim$ Binomial(16, $\theta$). After the study was completed, 13 of the 16 volunteers preferred the new flavour. What are the posterior probabilities for $H_0$ and $H_1$? And what is $BF_{01}$?

To begin, note that $\Pr(H_0|y)$ is the same as $\Pr(\theta \geq 0.6|y)$. We know that the Beta distribution is conjugate for the Binomial distribution and the posterior is Beta($\alpha + y$, $\beta + n - y$), or in this case, Beta(5.056+13, 5.056+16-13) = Beta(18.056, 8.056). Therefore:

$$\Pr(H_0|y = 13) = \int_{0.6}^{1} \frac{1}{Be(18.056, 8.056)} \theta^{18.056-1}(1-\theta)^{8.056-1} d\theta = 0.8448$$
$$\Pr(H_1|y = 13) = 1 - \Pr(H_0|13) = 0.1552$$

Note: the R code for $\Pr(\theta \geq 0.6)$ = `1-pbeta(0.6,18.056,8.056)` = 0.8447625. And the Bayes Factor:

$$BF_{01} = \frac{\Pr(H_0|y = 13)/\Pr(H_1|y = 13)}{\Pr(H_0)/\Pr(H_1)} = \frac{0.8448/0.1552}{0.265/0.735} = 15.1$$

which is between 3 and 20, thus "positive evidence" for $H_0$.

To evaluate the sensitivity of the resulting posterior probabilities and the Bayes Factor, three other priors for $\theta$ were considered: Beta(0.5,0.5) or Jeffreys' prior, Beta(1,1) or a Uniform(0,1), and Beta(2,2). The four prior densities are shown in Figure 1.

The posterior quantiles and means for $\theta$, given $y$=13 for $n$=16, as well as $\Pr(H_0)$ and $BF_{01}$ are shown in Table 1. The resulting posterior distributions are shown in Figure 2. As can be seen, the initial prior is the most skeptical regarding the preference for the new hummus.

Table 1: Numerical summaries of posterior quantities for taste preference study.

| $\pi(\theta)$ | $p_0$ | 0.25 | 0.50 | 0.75 | Mean | Pr(Ho\|y) | BF(0,1) |
|---|---|---|---|---|---|---|---|
| Beta(5.056, 5.056) | 0.265 | 0.63 | 0.70 | 0.76 | 0.69 | 0.84 | 15.06 |
| Beta(0.5, 0.5) | 0.436 | 0.74 | 0.81 | 0.87 | 0.79 | 0.96 | 34.43 |
| Beta(1, 1) | 0.400 | 0.72 | 0.79 | 0.85 | 0.78 | 0.95 | 30.81 |
| Beta(2, 2) | 0.352 | 0.69 | 0.76 | 0.82 | 0.75 | 0.93 | 24.60 |

[5]In R: 1-pbeta(0.6,5.056,5.056)=0.265.

Figure 1: Four prior distributions for $\theta$ in the hummus taste preference study. The vertical line at 0.6 marks the division between $H_0$ and $H_1$.

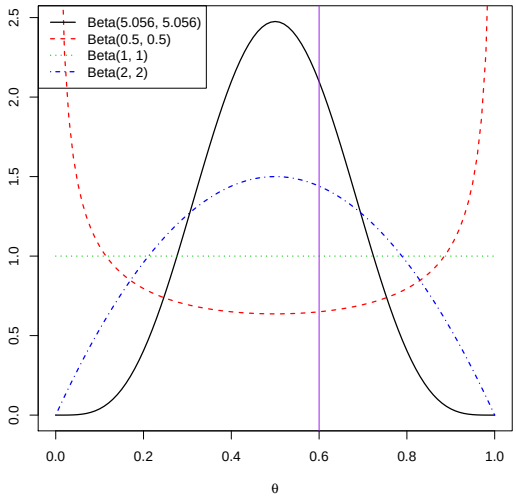
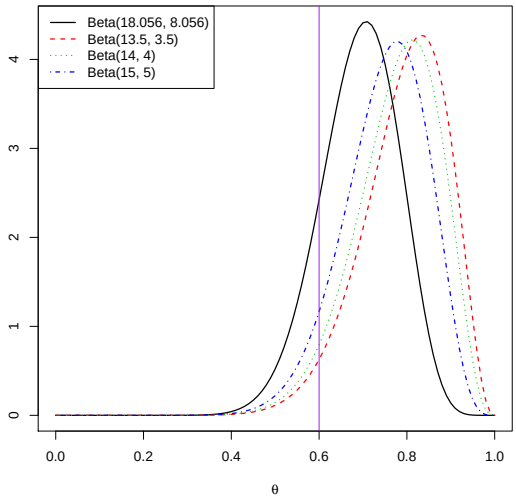



Figure 2: Four posterior distributions for $\theta$ in the hummus taste preference study given $y$=13 in $n$=16 trials. The vertical line at 0.6 marks the division between $H_0$ and $H_1$.



# 6 Example F: Simple Null and Composite Alternative

The sampling distribution is Poisson($\theta$). The null hypothesis is $H_0 : \theta = 5$ and the alternative is $H_1 : \theta \neq 5$, where $p_0$=0.7. A Gamma prior distribution is chosen for $\theta$ such that $E[\theta]$=5 with a CV of 0.1, thus a Gamma(100,20).

A random sample of $n$=8 is drawn yielding the following values

$$3, 3, 3, 3, 5, 7, 7, 4$$

Note: $\bar{y}$=4.375, and $\theta|\mathbf{y} \sim$ Gamma(100+$\sum_{i=1}^{8} y_i$, 20+$n$) = Gamma(135,28).

To find the posterior probabilities:

$$\begin{aligned}
\Pr(H_0|\mathbf{y}) &= \frac{\Pr(H_0, \mathbf{y})}{m(\mathbf{y})} \propto p_0 \prod_{i=1}^{8} \frac{e^{-5} 5^{y_i}}{y_i!} = 0.7 * \frac{e^{-40} 5^{35}}{\prod_{i=1}^{8} y_i!} = 9.12873e-08 \\
\Pr(H_1|\mathbf{y}) &= \frac{\Pr(H_1, \mathbf{y})}{m(\mathbf{y})} \propto p_1 \int_0^{\infty} \frac{e^{-\theta*8} \theta^{35}}{\prod_{i=1}^{8} y_i!} \frac{20^{100}}{\Gamma(100)} \theta^{100-1} e^{-20*\theta} \mathrm{d}\theta \\
&= 0.3 * \frac{1}{\prod_{i=1}^{8} y_i!} \frac{20^{100}}{\Gamma(100)} \frac{\Gamma(135)}{(28)^{135}} = 3.684845e-08
\end{aligned}$$

Then

$$\begin{aligned}
\Pr(H_0|\mathbf{y}) &= \frac{9.12873e-08}{9.12873e-08 + 3.684845e-08} = 0.7124 \\
\Pr(H_1|\mathbf{y}) &= \frac{3.684845e-08}{9.12873e-08 + 3.684845e-08} = 0.2876
\end{aligned}$$

And the Bayes Factor[6] for $H_0$ against $H_1$:

$$BF_{01} = \frac{0.7124/0.2376}{0.7/0.3} = 1.0617$$

which implies no evidence of $H_0$ over $H_1$ or vice versa.

# 7 Multiple Hypotheses

As said previously, multiple models can be viewed as multiple hypotheses. From an example by Lavine[7], a primary (elementary) school in Fresno, California had two high-voltage transmission lines nearby and the cancer rate amongst staff was a concern as 8 of the 145 staff had developed invasive cancers. Assume independence between staff and identical probabilities for cancer. Let $y$ denote the number developing cancer and $\theta$ the probability of cancer. Then $y \sim$ Binomial($n$=145, $\theta$) is the sampling model.

Based on data collected at a national level (for approximately the same age of the staff, mostly women, and number of years of working), the expected number of cancers for 145 staff was estimated to be 4.2. Translating that into a probability, one hypothesis was that $\theta$=4.2/145 $\approx$ 0.03. However, different individuals thought the rate was higher and three alternative hypotheses were postulated:

$$H_1 : \theta = 0.03, \quad H_2 : \theta = 0.04, \quad H_3 : \theta = 0.05, \quad H_4 : \theta = 0.06$$

[6] Note: a simpler calculation based on the right-most term in Eq'n 5 is $f(\mathbf{y}|H_0)/m(y)$, where $f(\mathbf{y}|H_0)$=$\frac{e^{-40} 5^{35}}{\prod_{i=1}^{8} y_i!}$= 1.304104e-07 and $m(y) = \frac{1}{\prod_{i=1}^{8} y_i!} \frac{20^{100}}{\Gamma(100)} \frac{\Gamma(135)}{(28)^{135}}$=1.228282e-07, and $BF_{01}$=1.0617.

[7] "What is Bayesian statistics and why everything else is wrong".

These four hypotheses can be viewed as 4 models. Lavine proposed that *a priori*, $H_1$ was as likely to be right as it was to be wrong, thus the prior for $H_1$ was $\Pr(H_1) = 1/2$. Then he assumed that any of the remaining hypotheses was equally likely, thus $\Pr(H_2) = \Pr(H_3) = \Pr(H_4)$=1/6. The posterior probabilities for the four hypotheses can be viewed as the relative weight of evidence for the competing theories:

$$\begin{aligned}\Pr(H_1|y=8) &= \frac{\Pr(y=8|H_1)\Pr(H_1)}{\sum_{i=1}^{4}\Pr(y=8|H_i)\Pr(H_i)} \\ &= \frac{0.03^8 * 0.97^{137} * \frac{1}{2}}{0.03^8 * 0.97^{137} * \frac{1}{2} + 0.04^8 * 0.96^{137} * \frac{1}{6} + 0.05^8 * 0.95^{137} * \frac{1}{6} + 0.06^8 * 0.94^{137} * \frac{1}{6}} = 0.23\end{aligned}$$

Repeating for $H_2$, $H_3$, and $H_4$:

$$\Pr(H_1|y=8) = 0.23, \quad \Pr(H_2|y=8) = 0.21, \quad \Pr(H_3|y=8) = 0.28, \quad \Pr(H_4|y=8) = 0.28$$

Thus, one could conclude that given the data and the priors, each of the four hypotheses are about equally likely. Or that the weight of evidence for each model is about the same. The posterior odds that the cancer rate is higher than the national average, or the posterior odds of $H_2$ or $H_3$ or $H_4$ against $H_1$ is (0.21+0.28+0.28)/0.23 =3.3. Given that the prior odds of $H_2$ or $H_3$ or $H_4$ and $H_1$ are 1, this is also the Bayes Factor and by the Kass and Raftery criteria this is just above the "positive evidence" lower bound of 3.

**Contrast with Frequentist Approach.** Lavine also carried out the frequentist analysis $H_0 : \theta = 0.3$ against the alternative $H_1 : \theta > 0.3$. The P-value is the probability of observing an outcome equal to what was observed, 8 occurrences of cancer in 145 staff, and anything more extreme in the direction of $H_1$[8]:

$$\begin{aligned}\Pr(Y \geq 8|\theta = 0.3) &= \Pr(Y = 8|\theta = 0.3) + \Pr(Y = 9|\theta = 0.3) + \ldots + \Pr(Y = 145|\theta = 0.3) \\ &= 1 - \Pr(Y < 8|\theta = 0.3, n = 145) = 0.0717\end{aligned}$$

This would be considered "significant" evidence against $H_0$ if the cut-off was 0.10. However, as Lavine points out this P-value does not account for how well the other hypotheses explain the data, information about things that did not happen (e.g., there were Not 9, nor 10, nor 11, and so on incidences of cancer), and the Likelihood Principle is not obeyed.

[8]This can be calculated in R by `1-pbinom(7,size=145,prob=0.03)`.

# R code

Taste preference code.

Code to produce the prior pdf plots in Figure 1.

```
# 4 sets of priors for theta
prior.a.set <- c(5.056,0.5,1,2)
prior.b.set <- c(5.056,0.5,1,2)
n <- 16
y <- 13
post.a.set <- prior.a.set + y
post.b.set <- prior.b.set + n-y

#- plot of prior distributions
theta.seq <- seq(0,1,by=0.01)
my.lwd <- 1.5
plot(theta.seq,dbeta(theta.seq,prior.a.set[1],prior.b.set[1]),type="l",
     xlab=expression(theta),ylab="",col=1,lty=1,xlim=c(0,1),lwd=my.lwd)
for(j in 2:4) {
  lines(theta.seq,dbeta(theta.seq,prior.a.set[j],prior.b.set[j]),
        col=j,lty=j,lwd=my.lwd)
}
abline(v=0.6,col="purple")
legend("topleft",legend=paste0("Beta(",prior.a.set,", ",prior.b.set,")"),
       lty=1:4,col=1:4,lwd=my.lwd)
```

Calculation of posterior quantiles, mean, Pr($H_0|y$), and $BF_{01}$.

```
#--- Calculation of quantiles, mean, Pr(Ho) and BF(0,1)
out.mat <- matrix(data=NA,nrow=4,ncol=6)
dimnames(out.mat) <- list(paste0("Beta(",prior.a.set,", ",prior.b.set,")"),
                          c("0.25","0.50","0.75","Mean","Pr(Ho|y)","BF(0,1)"))
for(i in 1:4) {
  out.mat[i,1:3] <- qbeta(c(0.25,0.50,0.75),post.a.set[i],post.b.set[i])
  out.mat[i,"Mean"] <- post.a.set[i]/(post.a.set[i]+post.b.set[i])
  p.Ho.y <- 1-pbeta(0.6,post.a.set[i],post.b.set[i])
  p.Ho   <- 1-pbeta(0.6,prior.a.set[i],prior.b.set[i])
  out.mat[i,"Pr(Ho|y)"] <- p.Ho.y
  out.mat[i,"BF(0,1)"]  <- (p.Ho.y/(1-p.Ho.y))/(p.Ho/(1-p.Ho))
}
print(round(out.mat,3))
```

And the plots of the posterior distributions:

```
plot(theta.seq,dbeta(theta.seq,post.a.set[1],post.b.set[1]),type="l",
     xlab=expression(theta),ylab="",col=1,lty=1,xlim=c(0,1),lwd=my.lwd)
for(j in 2:4) {
  lines(theta.seq,dbeta(theta.seq,post.a.set[j],post.b.set[j]),
        col=j,lty=j,lwd=my.lwd)
}
abline(v=0.6,col="purple")
legend("topleft",legend=paste0("Beta(",post.a.set,", ",post.b.set,")"),
       lty=1:4,col=1:4,lwd=my.lwd)
```

October 18, 2020