A Bayesian approach to statistical testing
Part 1: Normal distributions

Phil Garner & Bastian Schnell

August 6, 2019
The approach that follows is mathematical, but you only really need four formulae:

1. **Bayes’s (or Bayes’ or Bayes) theorem**

   \[ f(\theta | x) = \frac{f(x | \theta) f(\theta)}{f(x)} = \frac{f(x | \theta) f(\theta)}{\int_{-\infty}^{\infty} f(x | \theta') f(\theta') \, d\theta'} \]

   Where \( f(\theta | x) \) is shorthand for \( f_{\Theta}(\Theta = \theta | X_1 = x_1, X_2 = x_2, \ldots) \)

2. **Normal (or Gaussian) distribution**

   \[ f(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(x_i - \mu)^2}{2\sigma^2} \right) \]

3. **Binomial distribution**

   \[ f(k | \theta, n) = \binom{n}{k} \theta^k (1 - \theta)^{n-k} \]

4. **Beta distribution**

   \[ f(\theta | \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1 - \theta)^{\beta-1} \]
Hypothesis testing
  Overview
  Single sample test
  One and two tailed tests

Normal assumption
  Known variance
  Unknown variance
  Two sample tests
Origins of statistical testing

The important ones anyway

- This all arose from brewing beer!
- William Gosset was working for Guinness
- He published under the pseudonym Student, 1908

Example questions
Of interest to brewers

▶ Is this beer good enough to sell to customers?
  – Should we spend money promoting it?
▶ Is this beer better than this other beer?
  – Did we succeed in improving the beer?
▶ Can people tell the difference between this beer and another?
  – Is it worth updating our portfolio?

▶ Each of these is a distinct question
▶ Each requires a different statistical test
Hypothesis testing

The intuitive approach

You have some model with, e.g., known mean and variance: $\mu, \sigma^2$

You have data with, e.g., the obvious sample mean and variance: 
$
\bar{x}, \frac{1}{n} \sum (\bar{x} - x_i)^2
$

How close are these two distributions?

Intuitively, the shared area under the curves is an indicator

Note: It doesn’t have to be normally distributed, but often is

- Illustrate with single parameter $\theta$

Figure: The intuitive approach to testing.
Hypothesis testing
The formal approach

- Given some data, $x$, and parameter (model), $\theta$, we can define two mutually exclusive hypotheses:
  - $H_0$ the null hypothesis that $x$ and $\theta$ are compatible
  - $H_1$ the alternative hypothesis that $x$ and $\theta$ are incompatible
- Quantifying the alternative hypothesis can be difficult
  - You need to know what the alternative is
  - Typically you don’t; there are loads of alternatives
- A simpler approach is to assume the null hypothesis and show that it is unlikely
The Bayesian way

- Under $H_0$, the two distributions are the same!
- We can say something about the parameter under $H_0$

$$f(\theta | x) = \frac{f(x | \theta) f(\theta)}{f(x)} \quad (1)$$

$$f(\theta | x) \propto f(x | \theta) f(\theta) \quad (2)$$

- We can be lazy with the normalisation since that first line just says “and normalise the numerator”

Figure: Graphical model for the generic case; shaded $\implies$ observed
Testing the parameter
How compatible is the known parameter given the data?

We know the model parameter under $H_0$ – i.e., $\theta = \theta_0$ – $\theta$ is a sample from random variable $\Theta$

$f (\Theta = \theta_0 | x)$ is a point on a PDF

A more meaningful metric is the probability mass from that point over all more pessimistic values – this is the p-value.

The p-value is not quite the area from the “intuitive” slide, but it’s not far off

We reject $H_0$ if the p-value is small – e.g., $p < 0.05$

Figure: The meaning of the p-value for a normal-like variate.
One and two tailed tests

Illustration

Figure: Does the p-value on the left make sense? If not, do the one on the right
One and two tailed tests

The maths

Given the illustration on the previous slide there are actually three possibilities

1. If a decrease is expected

\[ p = \mathbb{P} (\Theta \leq \theta_0 | x) = \int_{-\infty}^{\theta_0} f (\theta | x) \, d\theta \] (3)

2. If an increase is expected

\[ p = \mathbb{P} (\Theta \geq \theta_0 | x) = \int_{\theta_0}^{\infty} f (\theta | x) \, d\theta \] (4)

3. If we don’t know whether an increase or decrease is expected

\[ p = 2 \times \min \left[ \mathbb{P} (\Theta \leq \theta_0 | x), \mathbb{P} (\Theta \geq \theta_0 | x) \right] . \] (5)
Outline

Hypothesis testing
  Overview
  Single sample test
  One and two tailed tests

Normal assumption
  Known variance
  Unknown variance
  Two sample tests
Example 1: Testing beer

Is this beer good enough to sell to customers?

▶ How good is this batch of beer?
▶ Give some to 20 drunks and ask them what they think
▶ Get them to rate it on the CQS
▶ We don’t sell beer rated less than “Excellent”
  – So we want the rating to be better than 80% = 0.8
▶ Assume the opinions are normally distributed

Figure: “Testers”

Figure: Continuous quality scale (CQS)

If the data are normally distributed, $\theta$ is $\mu$, the mean.

- All also depends on the variance $\sigma^2$ – may or may not be known.
- The distribution of $\mu$ is then

$$f(\mu | x, \sigma^2) = \frac{f(x | \mu, \sigma^2) f(\mu | \sigma^2)}{f(x | \sigma^2)}$$ (6)

$$f(\mu | x, \sigma^2) \propto f(x | \mu, \sigma^2) f(\mu | \sigma^2)$$ (7)

**Figure:** Graphical model for the simple Gaussian case
For normally distributed data, the likelihood follows in a standard way:

\[
f(x | \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \tag{8}\]

Some algebra:

\[
\begin{align*}
\vdots \\
= \frac{1}{\sqrt{2\pi}^n\sigma^n} \exp\left(-\frac{(\bar{x} - \mu)^2 + s^2}{2\sigma^2/n}\right) \tag{9}
\end{align*}
\]

where

\[
s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{10}
\]

is the sample variance.
A test statistic

- The above is enough to define a test
  - Take the prior, \( f(\mu \mid \sigma^2) \propto 1 \) (flat)
  - Take \( \sigma^2 \) to be known
- So,

\[
f(\mu \mid x, \sigma^2) \propto \frac{1}{\sqrt{2\pi}^n \sigma^n} \exp\left(-\frac{(\bar{x} - \mu)^2 + s^2}{2\sigma^2/n}\right)
\]

\[
= \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left(-\frac{(\bar{x} - \mu)^2}{2\sigma^2/n}\right)
\]

- \( \mu \) is normally distributed given the data
- This means that the variable

\[
\bar{z} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}
\]

follows a standard normal distribution
- \( \mu_0 \) is the value of \( \mu \) under \( H_0 \)
- The resulting test is called a \( z \)-test
The \( z \)-test

The single sample \( z \)-test is appropriate when

1. You know the mean, \( \mu = \mu_0 \)
   This is not normally a constraint
   Very often, \( \mu_0 = 0 \)
2. You know the variance \( \sigma^2 = \sigma_0^2 \)
   This is often a constraint!
   You need calibrated equipment
   *Drunks are not calibrated*

The \( z \)-test can be appropriate when \( \mu \) and \( \sigma^2 \) are approximating some other distribution

\[
p = P(M \leq \mu_0 | \mathbf{x}, \sigma^2) \]
\[
= \int_{-\infty}^{\mu_0} f(\mu | x, \sigma^2) \, d\mu
\]
Drunks are not calibrated

- From the z-test, we need

\[ z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \]  

(12)

- In a “human tester” situation, we do not know the variance

- It is tempting to substitute the sample variance instead

\[ z_s = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \quad s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \]  

(13)

- Student’s insight was to show that \( z_s \) is no longer normally distributed
Unknown variance
Setting the scene

- It is often not reasonable to assume $\sigma^2$ is known
- The Bayesian approach is to marginalise over it
  – This requires a prior
- Mathematically

\[
 f (\mu | \mathbf{x}) \propto \int_0^\infty f (\mathbf{x} | \mu, \sigma^2) f (\mu | \sigma^2) f (\sigma^2) \, d\sigma^2
\]

(14)

- We already used the uninformative prior
  \[
  f (\mu | \sigma^2) \propto 1
  \]
- We can use an uninformative prior \( f (\sigma^2) \propto \sigma^{-2} \)

Figure: The prior on $\sigma^2$
The required integral is easy enough (the trick is to integrate w.r.t. $\sigma^{-2}$)

$$f (\mu \mid \mathbf{x}) \propto \int_{0}^{\infty} f (\mathbf{x} \mid \mu, \sigma^{-2}) f (\sigma^{-2}) \, d\sigma^{-2}$$ (15)

$$\propto \int_{0}^{\infty} \frac{1}{\sqrt{2\pi}^{n} \sigma^{n}} \exp \left( -\frac{(\bar{x} - \mu)^{2} + s^{2}}{2\sigma^{2}/n} \right) \sigma^{2} \, d\sigma^{-2}$$ (16)

$$\vdots$$

(Some algebra; integral is gamma form)

$$\vdots$$

$$\propto \left( 1 + \frac{(\bar{x} - \mu)^{2}}{s^{2}} \right)^{-n/2}$$ (17)
Unknown variance

Tidying things up

- Define the degrees of freedom \( \nu = n - 1 \)
- So,

\[
f(\mu \mid x) \propto \left(1 + \frac{(\bar{x} - \mu)^2}{\nu s^2 / \nu}\right)^{-\frac{\nu+1}{2}}
\]

(18)

- Compare to a standard t-distribution

\[
f(t \mid x) = \frac{1}{\sqrt{\nu} B \left(\frac{1}{2}, \frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}
\]

(19)

- So, \( \mu \) is t-distributed given the data
- This means that the variable

\[
t = \frac{\bar{x} - \mu_0}{s / \sqrt{\nu}}
\]

(20)

Follows a standard t-distribution with \( \nu = n - 1 \) degrees of freedom
- As before, \( \mu_0 \) is the value of \( \mu \) under \( H_0 \)

https://en.wikipedia.org/wiki/Student’s_t-distribution
Unknown variance

The t-test

The single sample t-test is appropriate when

1. You know the mean \( \mu = \mu_0 \)
   This is the same as for a z-test
2. You do not know the variance
   i.e., it’s less constrained than a z-test;
   much closer to many real world problems

You can use the t-test when the normal is an approximation, but the approximations usually only hold for large \( n \), so it’s very similar to a z-test

Figure: The (one-tailed) p-value
Example 2: Comparing beer

Is this beer better than this other one?

- Instead of going for better than 80%, compare the beer to
  - A different beer
  - The same beer from last week
- Our $\mu_0$ is now a different test

Figure: Two beers

Figure: Continuous quality scale (CQS)

Two sample tests

Definition

- So far we assumed known \( \mu \)
- Commonly, the mean is unknown – but we have a second set of data
- \( H_0 \) is then that the two sets of data came from the same distribution

\[
\begin{align*}
\mu_1 &= \mu_2 \quad (21) \\
\mu_1 - \mu_2 &= 0 \quad (22)
\end{align*}
\]

Figure: Graphical model for the two sample case
For two groups of samples, there are two means and two sets of data.

Start with the joint distribution:

\[
f(\mu_1, \mu_2 | x_1, x_2, \sigma^2) = f(\mu_1 | x_1, \sigma^2) f(\mu_2 | x_2, \sigma^2)
\]

where

\[
\frac{1}{n} = \frac{1}{n_1} + \frac{1}{n_2}
\]

and we used a weak prior as before in the change of variable.
Two sample z-test

Recall that for the z-test we just normalise the likelihood

\[ f(\mu | x_1, x_2, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left(-\frac{(\bar{x}_1 - \bar{x}_2 - \mu)^2}{2\sigma^2/n}\right) \] (25)

so

\[ z = \frac{\bar{x}_1 - \bar{x}_2 - \mu_0}{\sigma/\sqrt{n}} \] (26)

For a two sample test, \( \mu_0 = 0 \) under \( H_0 \)
Two sample t-test

- Taking weak priors again, the marginalisation would then yield

\[
f(\mu | \mathbf{x}_1, \mathbf{x}_2) \propto \left(1 + \frac{n(\bar{x}_1 - \bar{x}_2 - \mu)^2}{n_1 s_1^2 + n_2 s_2^2}\right)^{-\frac{1}{2}(n_1 + n_2 - 1)}
\]

which is a t-distribution with

\[
\begin{align*}
t &= \frac{\bar{x}_1 - \bar{x}_2 - \mu_0}{S / \sqrt{\nu}}, \\
S &= \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n}}, \\
\nu &= n_1 + n_2 - 2
\end{align*}
\]

- Again, \(\mu_0 = 0\) under \(H_0\)
Bayes vs. classical

- The derivations above are Bayesian – i.e., the inference is about $\mu$
- The classical approach is to infer something about $\bar{x}$
- Use of Jeffreys priors leads to the same answers!
- In the normal case, the answers are the same
- When we look at binomial cases, the answers differ a little