Tutorial 3 Hypothesis Testing and Bayesian Inference

Overview

So far we have focused on descriptive statistics – that is, characterizing a set of measurements by graphs, tables, means, histograms – without considering our prior knowledge. As an exception to this, we have introduced the concept of calibrating our experiment against known standards. In a calibration, we are essentially saying “our prior knowledge is so good that we should adjust our measurements to match”. What if we have some prior information that cannot be ignored, but has a substantial degree of uncertainty? There are many ways of comparing or blending prior and new information. Two important approaches are Hypothesis Testing and Bayesian Inference. In this tutorial you will use both in the same scenario, but to slightly different ends.

Objectives:

a) Set up a hypothesis about the size of an object and test it with data using the Gaussian distribution (z-test).

b) Use Bayes Theorem to combine prior information and test data to develop the most likely “posterior probabilities” for the objects size.

Pre-tutorial questions:

Taylor: 5.35, 5.37

Mech 305/306 2017W Quiz 1 Question 3 (posted in tutorial module).

Background

The two types of inference considered in this tutorial have much in common, but have different origins and use different notation – so we'll keep the discussion separated.

Hypothesis Testing[1]

Statistical measurements are typically made to investigate or check some particular characteristic of the measured items. For example, an inspector working in a bottling plant may wish to check that the labeled amount of soft drink, say 1000ml, is being put into each bottle. This can be done by doing a "hypothesis test". A hypothesis is some as yet unproven belief or expectation- in a sense it is the embodiment of our prior knowledge of the issue.

In the soft drink bottle example, the expectation is that the mean content equals 1000ml. In statistical language, we say that the "null" hypothesis H_0 is the expectation that the mean bottle content is 1000ml and the "alternative" hypothesis H_1 is that is the opposite. If our measurements are sufficiently convincing to cause us to conclude that the mean bottle content is not 1000ml, then we "reject" the null hypothesis (and likely do something to fix the problem in the bottling plant). Otherwise, using the standard statistical language, we "fail to reject" the null hypothesis. This awkward double-negative language is used because we have not "proved" that the null hypothesis is "true", we have just not found sufficient cause to disagree. If we do disagree (i.e., reject the null hypothesis), the *p*-value tells us our chance of being wrong. In statistical language, there is a probability p that the null hypothesis is true AND we see the measurements that we do. Alternatively, if we agree (i.e., fail to reject the null hypothesis), the α-value indicates the chance that the test can give a false conclusion. In statistical language, our test has a probability α that it will fail to reject the null hypothesis when it should have done. This is called a Type I error.

[1] Hypothesis testing is briefly mentioned in Section 5.8 of the Taylor textbook, but not in great detail. There are many information sources available online, most rather mathematical. A simple but somewhat long video explaining hypothesis testing and Types I and II error is available at https://www.youtube.com/watch?v=k80pME7mWRM

Bayesian Inference

We are used to the idea that if have a device that "measures something", then the reading from that device is our "best estimate" of the quantity measured. This is very often not correct: often we have prior information that should be considered along with the measurements. The theorem postulated by Bayes (1701-1761) and further developed and published by Price provides a way to do this systematically. Bayesian inference is finding increasing use as a complement to "Big Data" and "Machine Learning". Recall Bayes Theorem,

$$P\langle A|B\rangle = \frac{P\langle B|A\rangle P\langle A\rangle}{P\langle B\rangle}$$

The Taylor text does not cover Bayes, but Wikipedia has a good section on this[2]. As written above, it is a pretty dry consequence of computing P(A and B) via conditional probabilities using "two routes". However, this becomes experimentally and philosophically more interesting when we interpret event B as the "evidence", such as a vector of measurements **X**, while event A is a hypothesis about a parameter of interest..

$$P\langle \theta|X\rangle = \frac{P\langle X|\theta\rangle P\langle \theta\rangle}{P\langle X\rangle}$$

Bayes Theorem can then be read as follows:

The "posterior" probability of θ having some particular true value, given observations X is equal to our "prior" estimate of P(θ), modified by the "evidence"[3],

$$\frac{P\langle X|\theta\rangle}{P\langle X\rangle}$$

In other words, we go into the experiment with some prior belief or hypothesis, and based on the observations, we update our views. For example, prior to getting a cancer test, our expectation of getting cancer is just the average rate (for our demographic group). After getting a screening test

[2] https://en.wikipedia.org/wiki/Bayes%27_theorem. A really nice video on Bayes is given at https://www.youtube.com/watch?v=R13BD8qKeTg&t=159s

[3] The exact definition of "the evidence" varies among sources, but this is a difference of words, not the mathematics. Often the denominator is called the "marginal likelihood" or "the evidence"

(observations X= positive or negative for cancer), we update our belief. If we had no prior knowledge of cancer rates, or if the test is “perfect”, then our posterior belief would simply be the test result (positive or negative). However, cancer is rare, and tests can give false positives, so our prior knowledge should count for something.

How do we apply Bayes to a more “engineering-like” example? Rather than present the general formulas, we’ll set up the equations as the Tutorial Assignment Part 2 is introduced, leaving you to solve them numerically.

Tutorial Assignment

The questions and file uploads will be on Canvas, but the assignments are explained below.

Part 1 Hypothesis Testing

The labels on the bins that contain bolts in the Mech stores were temporarily removed so that they could be repaired. You take a bolt from one of the unmarked bins and seek to identify it by measuring its diameter four times using a plastic ruler. The diameter measurements are 5.5, 5.0, 4.9, 5.6mm. Your experience with using a plastic ruler is that the standard deviation of bolt diameter measurements is typically 0.6mm, with Gaussian distribution.

You know that the metric bolts are stocked in integer mm diameters, e.g., 4, 5, 6, 7mm. The average of your four measurements is 5.25mm and so you form the opinion (hypothesis) that the actual bolt diameter is 5mm because it is the nearest integer.

Specific Deliverables for Part 1

1. Let's admit it, the test method is pretty crude. What is the likelihood (alpha value) that the measurements from your test method could suggest a different size even in the case when the bolt diameter actually was 5mm. (This is called a Type I error).
2. A friend comes by and insists that your bolt actually has diameter = 6mm. Based on your measurements you disagree, but because of measurement noise it is possible that you may be wrong. What is the level of your risk (P-value) that you incorrectly disagreed?

Part 2 Bayes Theorem

On further thought you realize that you can improve your diameter evaluation by getting some more information about the bolt supply[4]. You phone Markus in the shop, and he said that he had counted the bolts before removing the labels and the following quantities of the various diameters were present:

Bolt Diameter (mm):	*3*	*4*	*5*	*6*	*7*	*8*	*10*
Number in Stock:	*200*	*10*	*6*	*200*	*15*	*100*	*200*

What size bolt do you have, most likely? Before diving into the math below, think about the numbers here and use some logic and intuition to answer the question.

Mathematical Formulation

We want to evaluate the "posterior probability" $P(\theta_i|\mathbf{X})$ where θ_i are the alternate hypotheses that the bolt diameter is 3, 4, 5, 6, 7, 8, 10 mm *given the measurement vector* $\mathbf{X}$=[5.5, 5, 4.9, 5.6] *and the prior information about the frequency of bolts in Mech Stores.*

Let's start by considering only a few terms in the sums we will soon get. Suppose you have only one measurement x=5 mm and there were only two types of bolts in the storeroom: 6 5-mm bolts and 200 6-mm bolts. Our prior information is

$P(\theta_1)$=P(5mm)=6/206=0.029

$P(\theta_2)$=P(6mm)=0.971

The probability of getting a particular measurement, given that the true bolt size is θ is $P(x|\theta)$. The measurements should be continuously distributed, so we really mean $x \pm dx/2$

$$P(x|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right) dx$$

For our case, we have two calculations for the two alternative hypotheses:

[4] You should probably get some really good calipers to make the measurements, but you are keen to apply Baysian inference to this problem because you have heard that it is cool.

$$P(x = 5|\theta = 5) = \frac{1}{\sqrt{2\pi 0.6^2}} exp\left(-\frac{(5-5)^2}{2\ \ 0.6^2}\right) dx = 0.6649\ dx$$

$$P(x = 5|\theta = 6) = \frac{1}{\sqrt{2\pi 0.6^2}} exp\left(-\frac{(5-6)^2}{2\ \ 0.6^2}\right) dx = 0.1657\ dx$$

Now we can apply Bayes to get

$$P(\theta = 5|x = 5) = \frac{P(x = 5|\theta = 5)P(\theta = 5)}{P(x = 5|\theta = 6) + P(x = 5|\theta = 5)}$$

$$= \frac{(0.6649dx)(0.029)}{0.6649(0.029)dx + 0.1657\ (0.971)dx} = 0.107$$

$$P(\theta = 6|x = 5) = \frac{P(x = 5|\theta = 6)P(\theta = 6)}{P(x = 5|\theta = 6) + P(x = 5|\theta = 5)}$$

$$= \frac{(0.1657dx)(0.971)}{0.6649(0.029)dx + 0.1657\ (0.971)dx} = 0.893$$

Even though we measured the bolt to be 5mm, considering the prior information about the frequency of bolts AND our measurement accuracy, we should consider it more likely that the true bolt size is 6 mm.

Before adding the complexity of multiple measurements, note that

a) $\frac{dx}{\sqrt{2\pi\sigma}}$ cancels in both expressions as long as the measurement uncertainty is the same in all cases.

b) the denominator is the same for both hypotheses, so if we only want to calculate the ratios of probabilities, we can skip this step.

Now consider the case of having two measurements x_1=5 mm and x_2=4.5 mm with the same measurement uncertainty as before (σ=0.6). We'll simplify the calculation as suggested above to get

$$P(\theta = 5|x_1, x_2) \sim P(x_1, x_2|\theta = 5)P(\theta = 5)$$

Which should be compared with

$$P(\theta = 6|x_1, x_2) \sim P(x_1, x_2|\theta = 6)P(\theta = 6)$$

Here x_1, x_2 means that we have made independent measurements x_1 AND x_2. As usual, for a sequence of independent events, we get the probability as a product, *ie.*

$$P(\theta = 5|x_1, x_2) \sim P(x_1|\theta = 5)P(x_2|\theta = 5)P(\theta = 5)$$

And using only the parts of the Gaussian distribution that count,

$$P(\theta = 5|x_1, x_2) \sim exp\left(-\frac{(x_1 - 5)^2}{2\sigma^2}\right) exp\left(-\frac{(x_2 - 5)^2}{2\sigma^2}\right) P(\theta = 5)$$

Simplifying the exponentials,

$$P(\theta = 5|x_1, x_2) \sim exp\left(-\frac{(x_1-5)^2 + (x_2-5)^2}{2\sigma^2}\right) P(\theta = 5)\text{=0.021}$$

$$P(\theta = 6|x_1, x_2) \sim exp\left(-\frac{(x_1-6)^2 + (x_2-6)^2}{2\sigma^2}\right) P(\theta = 6)\text{=0.011}$$

Now it appears twice as likely (0.021/0.011) that the true bolt diameter is 5 mm. The extra evidence of x_2=4.5 mm changes the picture considerably.

Specific deliverables for Part 2

Write a matlab script to compute and plot on one graph

- the prior probabilities
- posterior probabilities
- the data **X**

The questions on Canvas are:

3. Make the graph for: priors as given above, σ=0.6 (base case), **X=**[5.5, 5, 4.9, 5.6]
4. Make the graph for: No prior information, data as given above, σ=0.6
5. Make the graph for: priors and data as given in question 3 but σ=0.1
6. Make the graph for: priors as given above, σ=0.1, but measurements are **X** X=[5.5,5,4.9,5.6,3.5,6.6,7,6.0]
7. Explain why the measurement uncertainty and prior information affects the calculated most likely bolt size.