

Lecture 7 Hypothesis Testing

1) Introduction:

The goal of hypothesis testing is to set up a procedure(s) to allow us to decide if a mathematical model ("theory") is acceptable in light of our experimental observations.

Examples:

Sometimes its easy to tell if the observations agree or disagree with the theory.

A certain theory says that Columbus will be destroyed by an earthquake in May 1992.

A certain theory says the sun goes around the earth.

A certain theory says that anti-particles (e.g. positron) should exist.

Often times its not obvious if the outcome of an experiment agrees or disagrees with the expectations.

A theory predicts that a proton should weigh 1.67×10^{-27} kg, you measure 1.65×10^{-27} kg.

A theory predicts that a material should become a superconductor at 30^0K , you measure the temperature of the transition to be 28^0 .

Often times we want to compare the outcomes of two experiments to check if they are consistent.

Experiment 1 measures the proton mass to be 1.67×10^{-27} kg, experiment 2 measures 1.62×10^{-27} kg

2) Types of Tests:

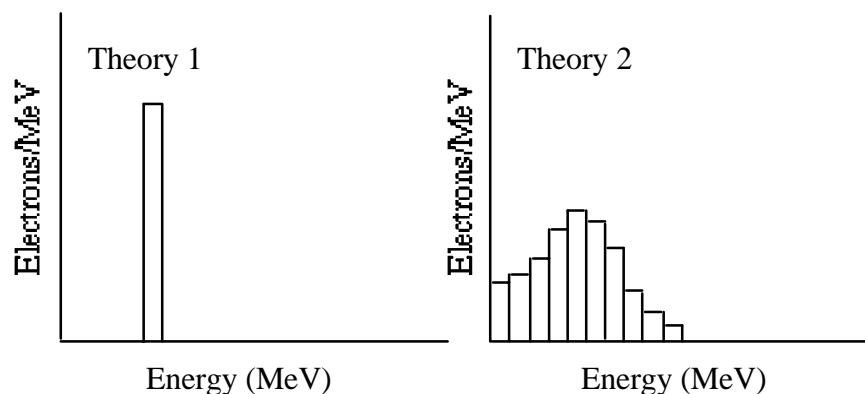
a) *Parametric Tests:* This type of test compares the values of parameters.

Example: does the mass of the proton = mass of the electron?

b) *Non-Parametric Tests:* These tests usually deal with the "shapes" of distributions.

Example: Consider the decay of a neutron.

Suppose we have two theories that predict the energy spectrum of the electron emitted in the decay of the neutron (beta decay):



Here a parametric test might not be sufficient to distinguish between the two theories. Both of these theories might predict the same average energy of the electron emitted in the decay of a neutron. However the shapes of their energy spectrums are quite different! Theory 1 is the spectrum expected if the neutron decays into two particles (e.g. proton + electron). Theory 2 is the spectrum expected if the neutron decays into 3 particles (proton + electron + ??). We would like a test that uses our data to differentiate between these two theories.

3) Confidence Levels and Intervals:

An informal definition of a confidence level (CL) is:

$$CL = 100 \times [\text{probability of the event happening by chance}]$$

The 100 in the above formula comes about because CL's are expressed as a percent (%).

Formally we can write for a continuous probability distribution P :

$$CL = 100 \times \text{prob}(x_1 \leq X \leq x_2) = 100 \times \int_{x_1}^{x_2} P(x) dx$$

Example: Suppose we measure some quantity (X) and we know that X is described by a Gaussian distribution with mean = 0 and standard deviation $\sigma = 1$. What is the CL for measuring $X \cdot 2$ (2σ from the mean)?

For a Gaussian distribution the probability for a measurement to be 2 or more standard deviations away from the mean is $\cdot 0.025$. Therefore the CL for this problem is:

$$CL = 100 \times \text{prob}(X \geq 2)$$

$$\text{prob}(X \geq 2) = \int_2^{\infty} P(\mu, \sigma, x) dx = \int_2^{\infty} P(0, 1, x) dx = \frac{1}{\sqrt{2\pi}} \int_2^{\infty} e^{-\frac{x^2}{2}} dx$$

$$CL = 100 \times 0.025 = 2.5\%$$

Note: to do this problem we needed to know the underlying probability distribution P . If the underlying probability distribution was not Gaussian (e.g. binomial) we could wind up with a very different numerical value for the CL. If you don't know P you are out of luck!

A word of caution about the use and interpretation of the CL. One can easily abuse the CL. For example suppose we have a scale of known accuracy (Gaussian with $\sigma = 10$ gm) and we weigh something to be 20 gm. Is there really a 2.5% chance that our object really weighs < 0 gm?? We must make sure that the probability distribution is defined in the region where we are trying to extract information.

There are some very subtle issues raised when CL's are discussed. The interpretation of confidence levels goes to the heart of the definition of probability and statistics and can mean different things to you depending on whether you take a Classical view or a Bayesian view of probability.

Example: Suppose we measure a value of x for the mean of a Gaussian distribution with an unknown mean μ . Suppose we know the standard deviation (σ) of the distribution. It is tempting to say:

“The probability that μ lies in the interval $[x-2\sigma, x+2\sigma] = 95\%$ ”

However according to Classical probability this is a *meaningless* statement! By definition the mean (μ) is a constant, not a random variable, thus μ does not have a probability distribution associated with it! What we can say is that we will reject any value of μ that gives a probability of $\cdot 5\%$ of obtaining our (measured) value of x .

There are also useful things called *Confidence Intervals* (CI). Here we know the CL but we want to know the value of x_1 and x_2 that give the CL.

Example: Suppose we have a Gaussian distribution with mean = 3 and $\sigma = 1$. What is the 68% CI for an observation?

We need to find the limits of the integral $[x_1, x_2]$ that satisfy:

$$0.68 = \int_{x_1}^{x_2} P(x) dx$$

For this problem $P(x)$ is given by the Gaussian distribution. In general one uses a table to find $[x_1, x_2]$. However, we can solve this problem if we remember that for a Gaussian distribution the area enclosed by $\pm 1\sigma$ is 0.68. Thus $x_1 = \mu - 1\sigma = 2$ and $x_2 = \mu + 1\sigma = 4$. Thus the CI is [2,4].

Example: Suppose an experiment observed no event, what is the 90% CL upper limit on the expected number of events?

$$CL = 0.90 = \sum_{n=1}^{\infty} \frac{e^{-\lambda} \lambda^n}{n!}$$

$$0.10 = \sum_{n=0}^{\infty} \frac{e^{-\lambda} \lambda^n}{n!} = e^{-\lambda}$$

$$\lambda = 2.3$$

i.e. if the expected number of events is greater than 2.3 events, then the probability of observing one or more events is greater than 90%.

Example: Suppose an experiment observed one event, what is the 95% CL upper limit on the expected number of events?

$$CL = 0.95 = \sum_{n=2}^{\infty} \frac{e^{-\lambda} \lambda^n}{n!}$$

$$0.05 = \sum_{n=0}^1 \frac{e^{-\lambda} \lambda^n}{n!} = e^{-\lambda} + \lambda e^{-\lambda}$$

$$\lambda = 4.74$$

4) Using Hypothesis Testing:

A procedure for testing a hypothesis is as follows:

- a) Measure something.
- b) Get a hypothesis (sometimes a theory) to test against your measurement.
- c) Calculate the CL that the measurement is from the theory.
- d) Accept or reject the hypothesis (or measurement) depending on some minimum acceptable CL.

One problem with this method has to do with d). How do we decide what is acceptable?

Example: How would we test the hypothesis: *the space shuttle is safe?*

What is an acceptable definition of safe?

One explosion per 10 launches?

One explosion per 1000 launches??

5) Hypothesis testing for Gaussian variables:

The following charts summarize some the different tests involving the mean and variance of Gaussian distributions.

If we want to test whether the mean of some quantity we have measured (\bar{x} = average from n measurements) is consistent with a known mean (μ_0) we have the following two tests:

Test	Conditions	Test Statistic	Test Distribution
$\mu = \mu_0$	σ^2 known	$\frac{x - \mu_0}{\sigma / \sqrt{n}}$	Gaussian $\mu = 0, \sigma = 1$
$\mu = \mu_0$	σ^2 unknown	$\frac{x - \mu_0}{s / \sqrt{n}}$	$t(n - 1)$

In the above chart $t(n - 1)$ stands for the “t-distribution” with $n - 1$ degrees of freedom.

Example: Do free quarks exist? Quarks are nature's fundamental building blocks and are thought to have electric charge (q) of either $(1/3)e$ or $(2/3)e$ ($e =$ charge of electron). Suppose we do an experiment to look for $q = 1/3$ quarks.

We measure: $q = 0.90 \pm 0.2$ This gives μ and σ

Quark theory: $q = 0.33$ This is μ_0

We want to test the hypothesis $\mu = \mu_0$ when σ is known. Thus we use the first line in the table.

$$z = \frac{x - \mu_0}{\sigma / \sqrt{n}} = \frac{0.9 - 0.33}{0.2 / \sqrt{1}} = 2.85$$

We want to calculate the probability for getting a $z \cdot 2.85$, assuming a Gaussian distribution.

$$\text{prob}(z \geq 2.85) = \int_{2.85}^{\infty} P(\mu, \sigma, x) dx = \int_{2.85}^{\infty} P(0, 1, x) dx = \frac{1}{\sqrt{2\pi}} \int_{2.85}^{\infty} e^{-\frac{x^2}{2}} dx = 0.002$$

The CL here is just 0.2%! What we are saying here is that if we repeated our experiment 1000 times then the results of 2 of the experiments would measure a value $q < 0.9$ if the true mean was $q = 1/3$. This is not strong evidence for $q = 1/3$ quarks!

If instead of $q = 1/3$ quarks we tested for $q = 2/3$ what would we get for the CL?

Now we have $\mu = 0.9$ and $\sigma = 0.2$ as before but $\mu_0 = 2/3$. We now have $z = 1.17$ and $\text{prob}(z < 1.17) = 0.13$ and the CL = 13%. Now quarks are starting to get believable!

Consider another variation of $q = 1/3$ problem. Suppose we have 3 measurements of the charge q:

$$q_1 = 1.1, q_2 = 0.7, \text{ and } q_3 = 0.9$$

We don't know the variance beforehand so we must determine the variance from our data.

Thus we use the second test in the table.

$$\mu = (q_1 + q_2 + q_3)/3 = 0.9$$

$$s^2 = \frac{\sum_{i=1}^n (q_i - \mu)^2}{n - 1} = \frac{0.2^2 + (-0.2)^2 + 0}{2} = 0.04$$

and

$$z = \frac{x - \mu_0}{s / \sqrt{n}} = \frac{0.9 - 0.33}{0.2 / \sqrt{3}} = 4.94$$

In this problem z is described by Student's t-distribution.

Note: Student is the pseudonym of statistician W.S. Gosset who was employed by a famous English brewery.

Just like the Gaussian distribution, in order to evaluate the t-distribution one must resort to a look up table (see for example Table 7.2 of Barlow). In this problem we want $\text{prob}(z < 4.94)$ when $n -$

1 = 2. The probability of $z < 4.94$ is 0.02, which is about 10X greater than the first part of this example where we knew the variance ahead of time.

We can also consider the situation where we have several independent experiments that measure the same quantity. We do not know the true value of the quantity being measured. Here we wish to know if the experiments are consistent with each other.

Test	Conditions	Test Statistic	Test Distribution
$\mu_1 - \mu_2 = 0$	σ_1^2 and σ_2^2 known	$\frac{x_1 - x_2}{\sqrt{\sigma_1^2 / n + \sigma_2^2 / m}}$	Gaussian $\mu = 0, \sigma = 1$
$\mu_1 - \mu_2 = 0$	$\sigma_1^2 = \sigma_2^2 = \sigma^2$ unknown	$\frac{x_1 - x_2}{Q\sqrt{1/n + 1/m}}$ $Q^2 \equiv \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2}$	$t (n+m-2)$
$\mu_1 - \mu_2 = 0$	σ_1^2 and σ_2^2 unknown	$\frac{x_1 - x_2}{\sqrt{s_1^2 / n + s_2^2 / m}}$	approx. Gaussian $\mu = 0, \sigma = 1$

Example: Suppose we wish to compare the results of two independent experiments to see if they agree with each other.

Exp 1 1.00 ± 0.01

Exp 2 1.04 ± 0.02

In this example we can use the first line of the table and set $n = m = 1$.

$$z = \frac{x_1 - x_2}{\sqrt{\sigma_1^2 / n + \sigma_2^2 / m}} = \frac{1.04 - 1.00}{\sqrt{(0.01)^2 + (0.02)^2}} = 1.79$$

Here z is distributed according to a Gaussian with $\mu = 0, \sigma = 1$. The probability for the two experiments to disagree by 0.04 (we don't care which experiment has the larger result so we use $|z|$) is:

$$prob(|z| \geq 1.79) = 1 - \int_{-1.79}^{1.79} P(\mu, \sigma, x) dx = 1 - \int_{-1.79}^{1.79} P(0, 1, x) dx = 1 - \frac{1}{\sqrt{2\pi}} \int_{-1.79}^{1.79} e^{-\frac{x^2}{2}} dx = 0.07$$

Thus 7% of the time we should expect the experiments to disagree at this level.

Is this acceptable agreement?

Basically: This is a personal choice.