Lecture 8
Hypothesis Testing

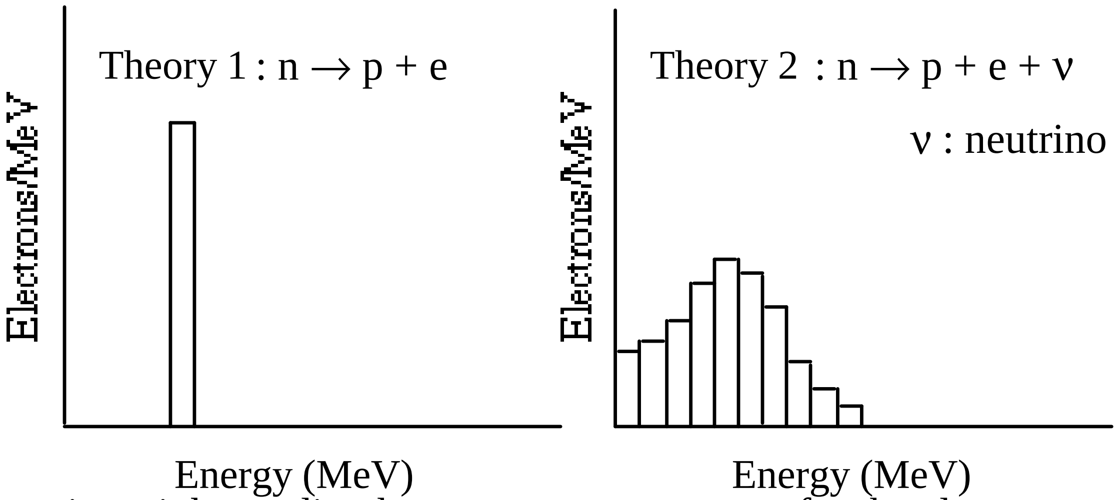
Taylor Ch. 6 and 10.6

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 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 300K, you measure 280K.
  - Often we want to compare the outcomes of two experiments to check if they are consistent.
    - Experiment 1 measures proton mass to be 1.67×10^{-27} kg, experiment 2 measures 1.62×10^{-27} kg.

Types of Tests

- *Parametric Tests*: compare the values of parameters.
  - Example: Does the mass of the proton = mass of the electron?
- *Non-Parametric Tests*: compare 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):



- Both theories might predict the same average energy for the electron.
  - A parametric test might not be sufficient to distinguish between the two theories.
- The shapes of their energy spectrums are quite different:
  - Theory 1: the spectrum for a neutron decaying into two particles (e.g. p + e).
  - Theory 2: the spectrum for a neutron decaying into three particles (p + e + ν).
- We would like a test that uses our data to differentiate between these two theories.

◆ We can calculate the χ^2 of the distribution to see if our data was described by a certain theory:

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - f(x_i, a, b\ldots))^2}{\sigma_i^2}$$

- $(y_i \pm \sigma_i, x_i)$ are the data points (n of them)
- $f(x_i, a, b\ldots)$ is a function that relates x and y
- accept or reject the theory based on the probability of observing a χ^2 larger than the above calculated χ^2 for the number of degrees of freedom.
- Example: We measure a bunch of data points $(y_i \pm \sigma_i, x_i)$ and we believe there is a linear relationship between x and y.

  $y = a + bx$

- If the y's are described by a Gaussian PDF then minimizing the χ^2 function (or using LSQ or MLM method) gives an estimate for a and b.
- As an illustration, assume that we have 6 data points and since we extracted a and b from the data, we have 6 - 2 = 4 degrees of freedom (DOF). We further assume:

$$\chi^2 = \sum_{i=1}^{6} \frac{(y_i - (a + bx_i))^2}{\sigma_i^2} = 15$$

- What can we say about our hypothesis that the data are described by a straight line?
- Look up the probability of getting $\chi^2 \geq 15$ by "chance":

$$P(\chi \geq 15, 4) \approx 0.006$$

- only 6 of 1000 experiments would we expect to get this result ($\chi^2 \geq 15$) by "chance".
- Since this is such a small probability we could reject the above hypothesis or we could accept the hypothesis and rationalize it by saying that we were unlucky.
- It is up to you to decide at what probability level you will accept/reject the hypothesis.

Confidence Levels (CL)

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

  CL = 100 x [probability of the event happening by chance]
  - The 100 in the above formula allows CL's to be expressed as a percent (%).
- We can formally write for a continuous probability distribution P:

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

For a CL, we know $P(x)$, x_1, and x_2.

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

$$\text{CL} = 100 \times prob(X \geq 2) = 100 \times \frac{1}{\sigma\sqrt{2\pi}} \int_{2}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \frac{100}{\sqrt{2\pi}} \int_{2}^{\infty} e^{-\frac{x^2}{2}} dx = 2.5\%$$

- To do this problem we needed to know the underlying probability distribution *P*.
- If the probability distribution was not Gaussian (e.g. binomial) we could have a very different CL.
- If you don't know *P* you are out of luck!

- Interpretation of the CL can be easily abused.
  - Example: We have a scale of known accuracy (Gaussian with $\sigma = 10$ gm).
    - We weigh something to be 20 gm.
    - Is there really a 2.5% chance that our object really weighs ≤ 0 gm??
    - ☞ probability distribution must be defined in the region where we are trying to extract information.

Confidence Intervals (CI)

- For a given confidence level, confidence intervals are the range $[x_1, x_2]$ that gives the confidence level.
  - Confidence interval's are not always uniquely defined.
  - We usually seek the minimum or symmetric interval.

For a CI, we know $P(x)$ and CL and wish to determine x_1, and x_2.

- Example: Suppose we have a Gaussian distribution with $\mu = 3$ 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 a Gaussian distribution the area enclosed by $\pm 1\sigma$ is 0.68.

$$x_1 = \mu - 1\sigma = 2$$
$$x_2 = \mu + 1\sigma = 4$$

  - ☞ confidence interval is [2,4].

Upper/Lower Limits

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

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

$$1 - \mathrm{CL} = 0.10 = 1 - \sum_{n=1}^{\infty} \frac{e^{-\lambda}\lambda^n}{n!} = \sum_{n=0} \frac{e^{-\lambda}\lambda^n}{n!} = e^{-\lambda}$$

$$\lambda = 2.3$$

If λ = 2.3, then 10% of the time we expect to observe zero events even though there is nothing wrong with the experiment!

- If the expected number of events is greater than 2.3 events,
  - 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?

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

$$1 - \mathrm{CL} = 0.05 = 1 - \sum_{n=2}^{\infty} \frac{e^{-\lambda}\lambda^n}{n!} = \sum_{n=0}^{1} \frac{e^{-\lambda}\lambda^n}{n!} = e^{-\lambda} + \lambda e^{-\lambda}$$

$$\lambda = 4.74$$

Procedure for Hypothesis Testing

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.

- Problem: How do we decide what is acceptable CL?
  - Example: What is an acceptable definition that the space shuttle is safe?
    - One explosion per 10 launches or per 1000 launches or…?

Hypothesis Testing for Gaussian Variables

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

Test	Condition	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	$\frac{x-\mu_0}{s/\sqrt{n}}$	$t(n-1)$

  - s: standard deviation extracted from the n measurements.
  - $t(n-1)$: Student's "t-distribution" with $n-1$ degrees of freedom.
    - Student is the pseudonym of statistician W.S. Gosset who was employed by a famous English brewery.
- 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.
  - Measure: q = 0.90 ± 0.2 = $\mu \pm \sigma$
  - Quark theory: q = 0.33 = μ_0
  - Test the hypothesis $\mu = \mu_0$ when σ is known:
    - Use the first line in the table:

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

    - Assuming a Gaussian distribution, the probability for getting a $z \geq 2.85$,

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

- If we repeated our experiment 1000 times,
  - two experiments would measure a value $q \geq 0.9$ 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?
  - $\mu = 0.9$ and $\sigma = 0.2$ as before but $\mu_0 = 2/3$.
    - $z = 1.17$
    - $\text{prob}(z \geq 1.17) = 0.13$ and CL = 13%.
    - quarks are starting to get believable!

- Consider another variation of $q = 1/3$ problem. Suppose we have 3 measurements of the charge q: $q_1 = 1.1$, $q_2 = 0.7$, and $q_3 = 0.9$
  - We don't know the variance beforehand so we must determine the variance from our data.
    - use the second test in the table:

$$\mu = \tfrac{1}{3}(q_1 + q_2 + q_3) = 0.9$$

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

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

  - Table 7.2 of Barlow: $\text{prob}(z \geq 4.94) \approx 0.02$ for $n - 1 = 2$.
    - 10X greater than the first part of this example where we knew the variance ahead of time.

- 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.
  - We wish to know if the experiments are consistent with each other.

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

$$Q^2 \equiv \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2}$$

- Example: We compare 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

  - Use the first line of the table and set $n = m = 1$.

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

    - z is distributed according to a Gaussian with $\mu = 0,\ \sigma = 1$.
    - Probability for the two experiments to disagree by ≥ 0.04:

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

- We don't care which experiment has the larger result so we use $\pm z$.
  - 7% of the time we should expect the experiments to disagree at this level.
  - Is this acceptable agreement?