PHYS 500: Research Methodology

Lecture 2: Basic Principles of Theory of Errors

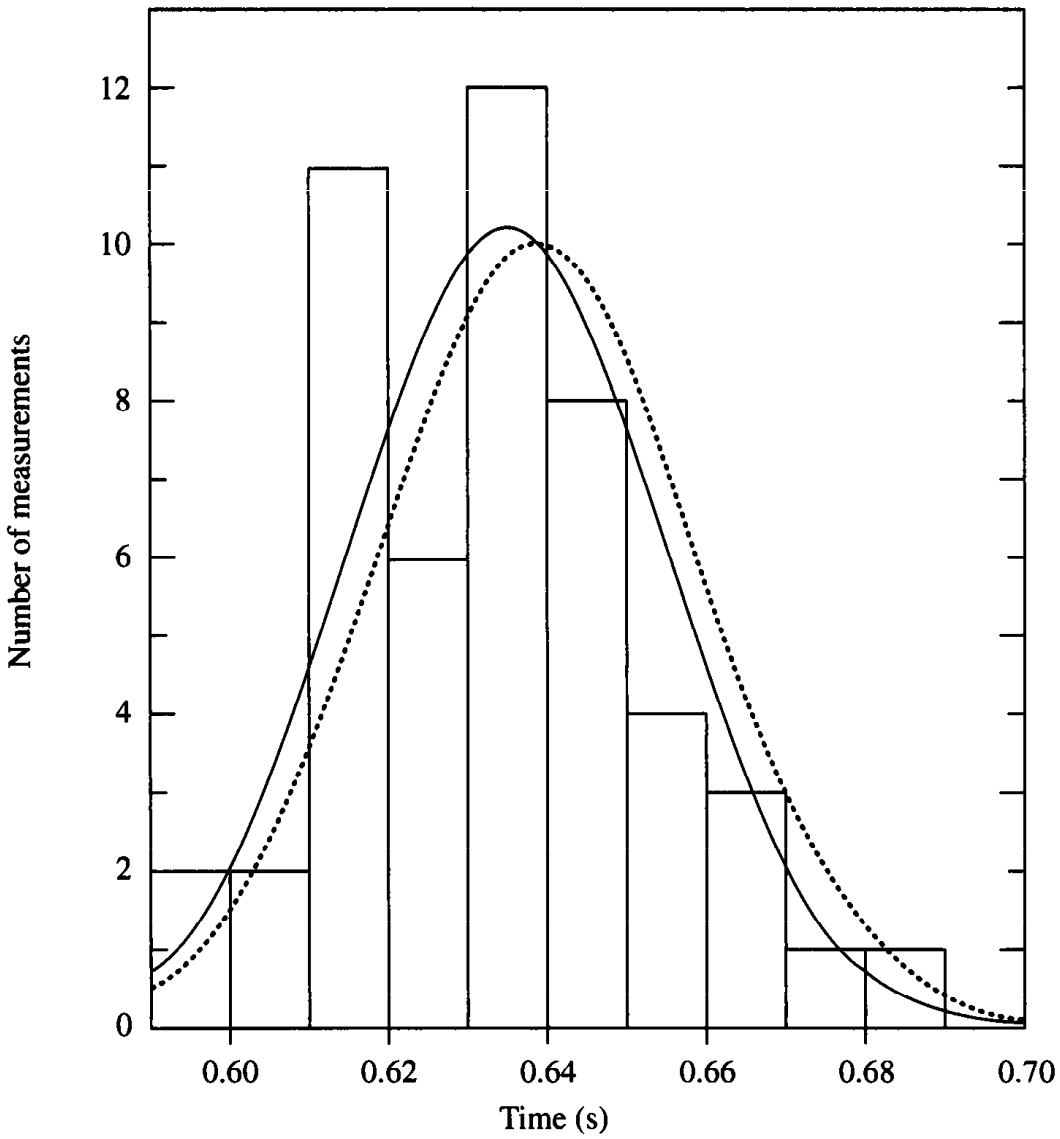
Parent and Sample Distributions

- When we make more and more measurements a pattern will emerge from the data. On the average we expect the data to be distributed around the correct value.
- If we could make an infinite number of measurements, then we could exactly determine the distribution of the data points. This is not possible in practice but we can hypothesize the existence of such a distribution that determines the probability of getting any particular observation in a single measurement. This is called **the parent distribution**.
- Our measurements make up the so called **sample distribution**.
- In the limit of infinite measurements the sample distribution becomes the parent distribution.

Parent and Sample Distributions

- In order to determine the parameters of the parent distribution, we assume that the results of experiments asymptotically approach the parent quantities as the number of measurements approaches infinity; that is, the parameters of the experimental distribution equal the parameters of the parent distribution **in the limit of an infinite number of measurements.**
- If we specify that there are *N* observations in a given experiment, then we can denote this by:

$$\left(\text{parent parameter}\right) = \lim_{N\to\infty}\left(\text{experimental parameter}\right)$$



Histogram of measurements of the time for a ball to fall 2.00 m. The solid Gaussian curve was calculated from the mean ($\bar{T}$ = 0.635 s) and standard deviation (s = 0.020 s) estimated from these measurements. The dashed curve was calculated from the parent distribution with mean μ = 0.639 s and standard deviation σ = 0.020 s.

Continuous and discrete data

- A random variable is any function of the data. The event space is the set of all possible values of a random variable. A random variable which can have any value between two arbitrarily given values in the event space is called a **continuous variable**;
- Conversely, if the variable can only have certain values it is called a discrete variable. In the same manner, data described by discrete or continuous variables are called **discrete data** or **continuous data** respectively.

Characteristic quantities

- The probability distributions are characterized by the following quantities:

1. The **mean value** $\overline{x}$ **(**sample) or μ (parent).
2. The **median** $\mu_{1/2}$ (of the parent distribution).
3. The **mode** or **most probable value** μ_{max}.
4. A parameter that is a measure of the observation from the average value is the **standard deviation** σ. This is related to the concept of the **variance** s^2(sample) or σ^2 (parent).

For a symmetrical distribution the above values 1, 2, 3 are the same.

- **Note:** The symbol σ is often used to represent the best estimate of the standard deviation of the parent distribution from a sample distribution.

The mean value (discrete data)

- The **mean value** $\overline{x}$ **(**sample) of a set of N numbers x_i is given by:

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- The mean of a function of x, is given by:

$$\left\langle f(x) \right\rangle = \frac{1}{N}\sum_{i=1}^{N} f\left(x_i\right)$$

- If the N data points are classified by a frequency distribution in m intervals, and if n_j stands for the number of entries in the interval j, then:

$$\left\langle f(x) \right\rangle = \frac{1}{N}\sum_{j=1}^{m} n_j x_j$$

The mean value (continuous data)

- The **mean value** $\overline{x}$ **(**sample) of a is given by:

$$\mu = \int_{-\infty}^{+\infty} xp(x)\,dx$$

- The mean of a function of the continuous data *x*, is given by:

$$\langle f(x) \rangle = \int_{-\infty}^{+\infty} f(x)p(x)\,dx$$

The median of a discrete set of data

- The **median** is the value separating the higher half of a data **sample**, a **population**, or a **probability distribution**, from the lower half. In simple terms, it may be thought of as the "middle" value of a data set. For example, in the data set {1, 3, 3, 6, 7, 8, 9}, the median is 6, the fourth number in the sample.
- The median of a finite list of numbers can be found by arranging all the numbers from smallest to greatest.
- If there are an even number of observations, then there is no single middle value; the median is then usually defined to be the **mean** of the two middle values. For example, in the data set: 1, 2, 3, 4, 5, 6, 8, 9. The median is the mean of the middle two numbers: this is (4 + 5) ÷ 2, which is 4.5.
- The formula used to find the middle number of a data set of *n* numbers is (n + 1) ÷ 2. This either gives the middle number (for an odd number of values) or the halfway point between the two middle values. For example, with 14 values, the formula will give 7.5, and the median will be taken by averaging the seventh and eighth values.

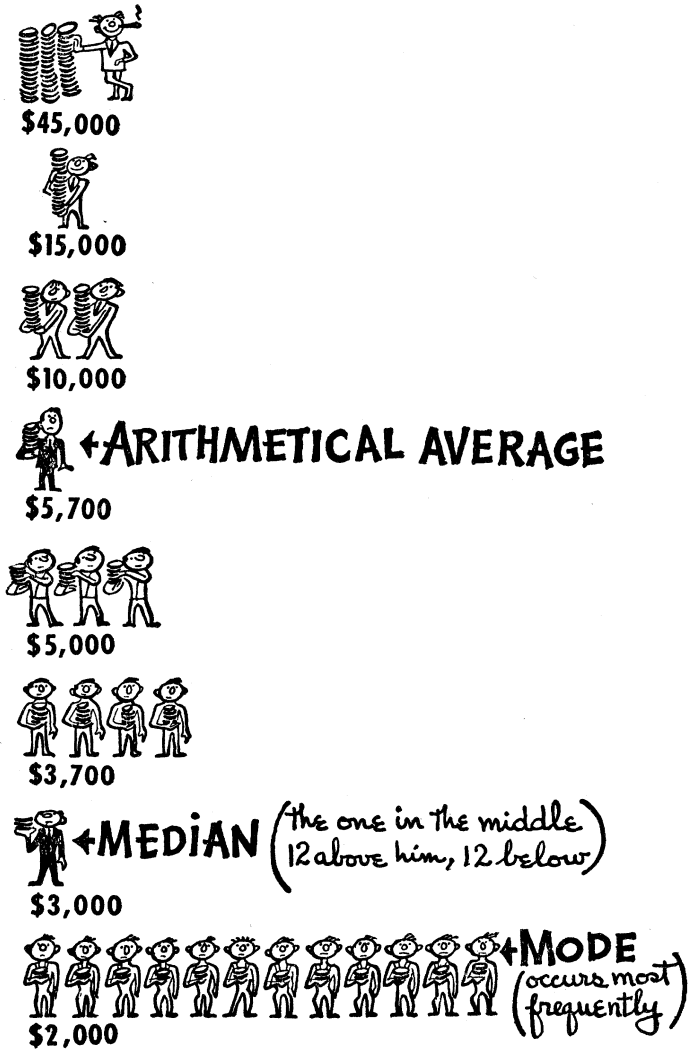
The median of the parent distribution

- The **median** $\mu_{1/2}$ is defined as the value for which, in the limit of infinite measurements, half the observations will be less than the median and half will be greater.

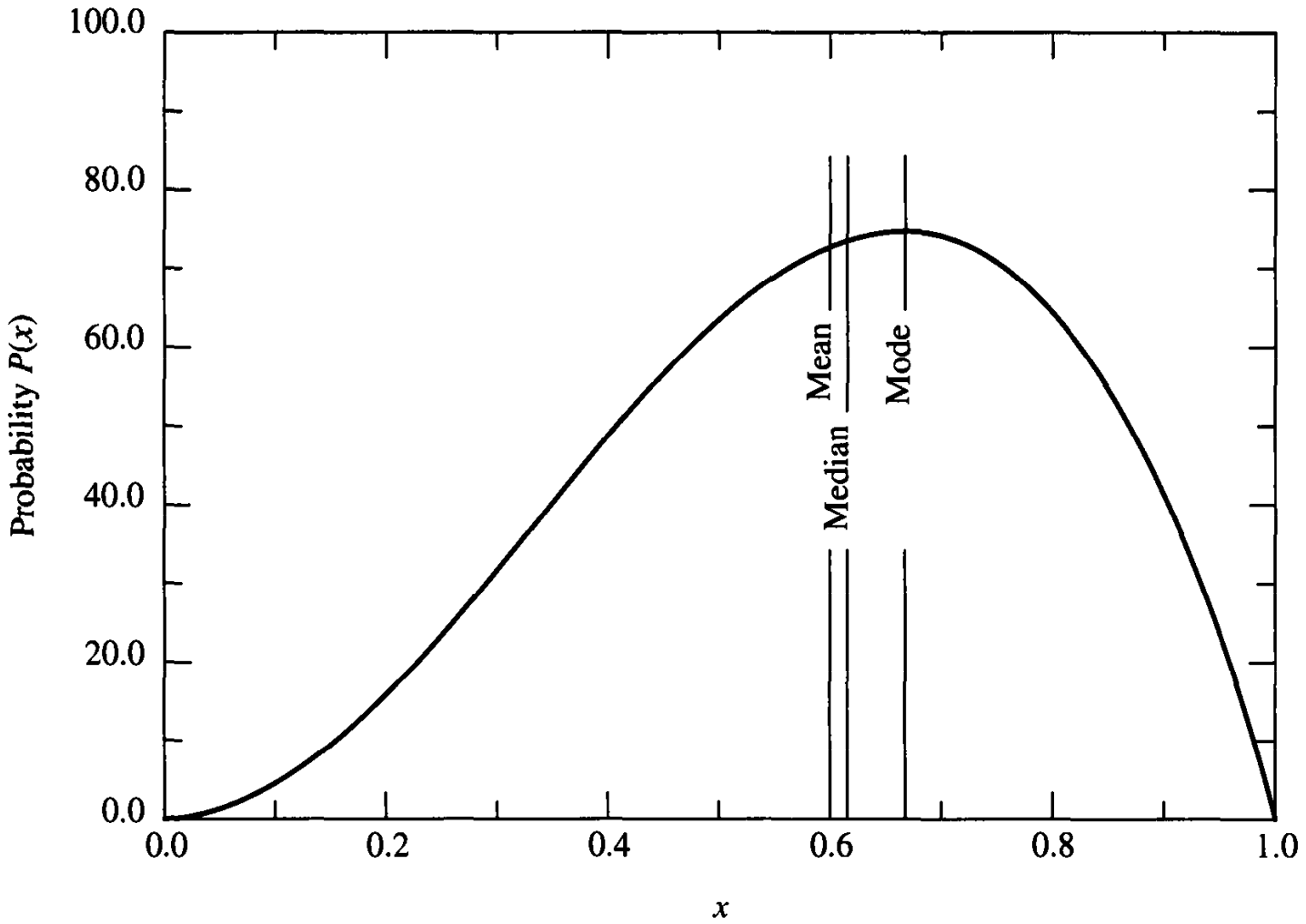
$$\int_{-\infty}^{x_{median}} p(x)\,dx = \int_{x_{median}}^{+\infty} p(x)\,dx = 0.5$$

The mode

- The **mode** or **most probable value** μ_{max}. Is the value for which the parent distribution has the greatest value.
- The mode is not necessarily unique: if a distribution has two maxima, we call it **bimodal**, if it has several maxima, we call it **multimodal**.
- When only one maximum is present the mode is also called most probable value.



The distribution of the monthly income of Americans around the year 1950. This pictorial representation explains well the differences between mean, mode and median. Which of the three describes the most important property? (Taken from [6])



Asymmetric distribution illustrating the positions of the mean, median, and mode of the variable.

Variance and Standard Deviation-a

- A parameter that is proper to measure the dispersion of the observations is the **standard** deviation. This is defined as as the square root of the **variance σ^2** given by:

$$\sigma^2 \equiv \lim_{N\to\infty}\left[\frac{1}{N}\sum_{i=1}^{N}\left(x_i-\mu\right)^2\right]=\lim_{N\to\infty}\left(\frac{1}{N}\sum_{i=1}^{N}x_i^2\right)-\mu^2$$

- The corresponding expression is for the **variance s^2 (or population standard deviation)** of a sample population is defined as:

$$s^2 \equiv \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i-\overline{x}\right)^2$$

Variance and Standard Deviation-b

- If the probability density function is a continuous smoothly varying function $p(x)$ of the observed value x, we replace the sum over the individual observations by an integral over all values of x multiplied by the probability $p(x)$. In this case we have:

$$\sigma^2 \equiv \int_{-\infty}^{\infty} \left(x-\mu\right)^2 p(x)\,dx = \int_{-\infty}^{\infty} x^2 p(x)\,dx - \mu^2$$

- It can be shown that the variance s^2 of values obtained by repeated measurements of a quantity remains almost constant, regardless of how many measurements are made.

Discussion-a

- The mean is the best estimate of the “true” value under the prevailing experimental conditions.
- The variance s^2 and the standard deviation s characterize the uncertainties associated with our experimental attempts to determine the “true” values. For a given number of observations, the uncertainty in determining the mean of the parent distribution is proportional to the standard deviation of that distribution.
- The standard deviation s is, therefore, an appropriate measure of the uncertainty due to fluctuations in the observations in our attempt to determine the “true” value

Discussion-b

$$a \equiv \lim_{N\to\infty}\left[\frac{1}{N}\sum_{i=1}^{N}\left|x_i-\mu\right|\right]$$

- Although, in general, the distribution, resulting from purely statistical errors can be described well by the two parameters: the mean and the standard deviation, we should be aware that, at distances a few standard deviations from the mean of an experimental distribution, non statistical errors may dominate.
- In specially severe cases, it may be preferable to describe the spread of the distribution in terms of the **average deviation** *a*, rather than the standard deviation, because the later tends to deemphasize measurements that are far from the mean.
- There are also distributions for which the variance does not exist. The average deviation or some other quantity must be used as a parameter to indicate the spread of the distribution in this case.

Discussion-c

- What is the connection between the probability distribution of the parent population and an experimental sample we obtain?

1. From our experimental data points we can determine a sample frequency distribution that describes the way in which these particular data point are distributed over the range of possible data points. The shape and magnitude of the sample distribution vary from sample to sample.
2. From the parameters of the sample probability distribution we can estimate the parameters of the probability distribution of the parent population of possible observations. Our best estimate for the mean μ is the mean of the sample distribution $\overline{x}$, and the best estimate for the variance σ^2 is the sample variance s^2. Even the shape of this parent distribution must be estimated or assumed.

Discussion-d

3. From the estimated parameters of the parent distribution we estimate the results sought. In general, we shall assume that the estimated parameters of the parent distribution are equivalent to the “true” values, but the estimated parent distribution is a function of the experimental conditions as well as the “true” values, and these may not necessarily be separable.

The Average Value-a

- Let's consider that we measure the same quantity x for N times and we get the recordings:

$$x_1,\ x_2,\ x_3,\ \ldots,\ x_N$$

- The average value of this quantity is defined as:

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- This is not the real value so we have to estimate the relevant error δx. This is the region in which the real value lies with a certain probability.

$$x = \overline{x} \pm \delta x$$

The Average Value-b

- The mathematical theory of errors says that if we wish the real value to be in the region $x = \bar{x} \pm \delta x$ with a probability 68% then the error must be given by the formula:

$$\delta x = \sqrt{\frac{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2}{N(N-1)}}$$

- This is the so-called **absolute error of the average value**.
- We see that as the number of repetitions *N* is increased the error decreases.

Hint: Be careful, we can find the average value only if we measure the same quantity and we expect to find the same value in each measurement!

The Relative Error

- To check whether or not the error in an experiment is large or small we use the concept of the **relative error.** This is defined as follows:

$$\eta = \frac{\delta x}{\overline{x}}$$

- This sometimes is expressed in a percentage form as:

$$\eta = \frac{\delta x}{\overline{x}} \cdot 100\%$$

- As we do for the error, we quote these quantities to 1 or 2 significant digits.

How we quote the value of a quantity

- **Step 1:** Calculate the average value
- **Step 2:** Calculate the uncertainty in the quantity, making clear the method used. Round the uncertainty to one significant figure (or two if the first non-zero figure is either “1” or “2”)
- **Step 3:** Quote the average and the uncertainty to the appropriate number of significant figures.
- **Step 4:** Include the units of the quantity