Intermediate Lab
PHYS 3870

Lecture 3

Distribution Functions

References: Taylor Ch. 5 (and Chs. 10 and 11 for Reference)
Taylor Ch. 6 and 7
Also refer to “Glossary of Important Terms in Error Analysis”
“Probability Cheat Sheet”
Intermediate Lab

PHYS 3870

Distribution Functions
Practical Methods to Calculate Mean and St. Deviation

We need to develop a good way to tally, display, and think about a collection of repeated measurements of the same quantity.

Here is where we are headed:
• Develop the notion of a probability distribution function, a distribution to describe the probable outcomes of a measurement
• Define what a distribution function is, and its properties
• Look at the properties of the most common distribution function, the Gaussian distribution for purely random events
• Introduce other probability distribution functions

We will develop the mathematical basis for:
• Mean
• Standard deviation
• Standard deviation of the mean (SDOM)
• Moments and expectation values
• Error propagation formulas
• Addition of errors in quadrature (for independent and random measurements)
• Schwartz inequality (i.e., the uncertainty principle) (next lecture)
• Numerical values for confidence limits (t-test)
• Principle of maximal likelihood
• Central limit theorem
Two Practical Exercises in Probabilities

Flip a penny 50 times and record the results.

Roll a pair of dice 50 times and record the results.

Grab a partner and a set of instructions and complete the exercise.
Two Practical Exercises in Probabilities

Flip a penny 50 times and record the results

Group Two Instructions
1. Flip penny 50 times
2. Record each result as “H” or “T” in list below

|  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

Group One Instructions
1. Flip penny 50 times
2. Tally results on list below

Heads:

Tails:

What is the asymmetry of the results?
Two Practical Exercises in Probabilities

Flip a penny 50 times and record the results

Group Two Instructions
1. Roll two dice 50 times
2. Record each result as “H” or “T” in list below

```
H  H  T  H  T  T  H  H  T  T
H  H  H  T  T  H  H  T  H  H
H  T  H  T  H  T  H  T  T  T
H  H  T  T  H  H  T  T  T  H
H  T  H  H  T  T  H  T  T  T
```

Group One Instructions
1. Flip penny 50 times
2. Tally results on list below

- Heads: \( \frac{54}{100} \) = 54%
- Tails: \( \frac{46}{100} \) = 46%

??% asymmetry

4% asymmetry

What is the asymmetry of the results?
**Two Practical Exercises in Probabilities**

Roll a pair of dice 50 times and record the results

<table>
<thead>
<tr>
<th>Group One Instructions</th>
<th>Group Two Instructions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roll two dice 50 times</td>
<td>Roll two dice 50 times</td>
</tr>
<tr>
<td>Record results on table below, checking one box for each die</td>
<td>Record each result in list below</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

What is the mean value? The standard deviation?
Two Practical Exercises in Probabilities

Roll a pair of dice 50 times and record the results

**Group Two Instructions**
Roll two dice 50 times
Record each result in list below

<table>
<thead>
<tr>
<th>3</th>
<th>4</th>
<th>10</th>
<th>6</th>
<th>9</th>
<th>7</th>
<th>9</th>
<th>3</th>
<th>11</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>6</td>
<td>9</td>
<td>7</td>
<td>11</td>
<td>12</td>
<td>6</td>
<td>11</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>8</td>
<td>4</td>
<td>11</td>
<td>10</td>
<td>12</td>
<td>11</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>12</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>7</td>
<td>10</td>
<td>9</td>
<td>11</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>10</td>
<td>5</td>
<td>8</td>
<td>7</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>7</td>
</tr>
</tbody>
</table>

What is the mean value?

The standard deviation?

What is the asymmetry (kurtosis)?

What is the probability of rolling a 4?

**Group One Instructions**
Roll two dice 50 times
Record results on table below, checking one box for each die

Mean = 7.3
St. Dev. = 2.8
Discrete Distribution Functions

A data set to play with

\[ 26, 24, 26, 28, 23, 24, 25, 24, 26, 25. \] \hspace{1cm} (5.1)

\[ 23, 24, 24, 24, 25, 25, 26, 26, 26, 28. \] \hspace{1cm} (5.2)

Written in terms of “occurrence” \( F \)

<table>
<thead>
<tr>
<th>( x_k )</th>
<th>( 22 )</th>
<th>( 23 )</th>
<th>( 24 )</th>
<th>( 25 )</th>
<th>( 26 )</th>
<th>( 27 )</th>
<th>( 28 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( F_k )</td>
<td>0.0</td>
<td>0.3</td>
<td>0.1</td>
<td>0.2</td>
<td>0.3</td>
<td>0.0</td>
<td>0.1</td>
</tr>
</tbody>
</table>

The mean value

\[
\bar{x} = \frac{\sum x_i}{N} = \frac{23 + 24 + 24 + 24 + 25 + \ldots + 28}{10}.
\]

This equation is the same as

\[
\bar{x} = \frac{23 + (24 \times 3) + (25 \times 2) + \ldots + 28}{10}
\]

or in general

\[
\sum_k n_k \cdot x_k = \frac{\sum_k (n_k \cdot x_k)}{N} = \frac{\sum_k (n_k \cdot x_k)}{\sum_k n_k}
\]

In terms of fractional expectations

Fractional expectations

\[ F_k = \frac{n_k}{N} \]

Normalization condition

\[ 1 = \sum_k (F_k) \]

Mean value

\[ \bar{x} = \sum_k (F_k \cdot x_k) \]

(This is just a weighted sum.)
Limit of Discrete Distribution Functions

"Normalizing" data sets

\[ f_k \Delta_k = \text{fraction of measurements in } k\text{th bin.} \]

\[ f_k \equiv \text{fractional occurrence} \]
\[ \Delta_k \equiv \text{bin width} \]

Mean value:

\[ X = \sum_k F_k x_k = \sum_k (f_k \Delta_k) x_k \]

Normalization:

\[ 1 = \frac{\sum_k (f_k \Delta_k) x_k}{\sum_k \Delta_k} \]

Expected value:

\[ \text{Prob}(4) = \frac{F_4}{N} = f_4 \]
Limits of Distribution Functions

Consider the limiting distribution function as N \( \rightarrow \infty \) and \( \Delta_k \rightarrow 0 \)

Larger data sets

Figure 5.3. Histogram for 100 measurements of the same quantity as in Figure 5.2.

Mathcad Games:

- # Data Pts: \( N = 6000 \)
- Mean: \( \mu = 10 \)
- Std. Dev.: \( \sigma = 5 \)

SDOM: \( \frac{\sigma}{\sqrt{N}} = 0.065 \)

Fractional Error: \( \left( \frac{\sigma}{\sqrt{N}} \right) \cdot \frac{1}{\mu} = 0.645\% \)
Continuous Distribution Functions

Meaning of Distribution Interval

Thus

\[ \int_{a}^{b} f(x) \, dx = \text{fraction of measurements that fall within } a < x < b \]

and by extension

Central (mean) Value

\[ \sum_{k} (F_{k} \cdot x_{k}) = X \Rightarrow \int_{-\infty}^{+\infty} x \, f(x) \, dx = \bar{x} = \langle x \rangle \] (5.15)

Normalization of Distribution

\[ \sum_{k} (F_{k}) = 1 \Rightarrow \int_{-\infty}^{+\infty} f(x) \, dx = 1 \]

Width of Distribution

\[ \sum_{k} [(x_{k} - \bar{x})^{2} \cdot F_{k}] = \sigma_{X}^{2} \Rightarrow \int_{-\infty}^{+\infty} (x - \bar{x})^{2} \, f(x) \, dx = \sigma_{X}^{2} = \langle (x - \bar{x})^{2} \rangle \] (5.16)
Moments of Distribution Functions

The first moment for a **probability distribution function** is

\[ \bar{x} \equiv \langle x \rangle = \text{first moment} = \int_{-\infty}^{+\infty} x \, f(x) \, dx \]

For a **general distribution function**, \n
\[ \bar{x} \equiv \langle x \rangle = \text{first moment} = \frac{\int_{-\infty}^{+\infty} x \, g(x) \, dx}{\int_{-\infty}^{+\infty} g(x) \, dx} \]

Generalizing, the \( n \)th moment is

\[ x_n \equiv \langle x^n \rangle = \text{nth moment} = \frac{\int_{-\infty}^{+\infty} x^n \, g(x) \, dx}{\int_{-\infty}^{+\infty} g(x) \, dx} = \int_{-\infty}^{+\infty} x^n \, f(x) \, dx \]

- 0th moment \( \equiv N \)
- 1st moment \( \equiv \bar{x} \)
- 2nd moment \( \equiv \langle (x - \bar{x})^2 \rangle \to \langle x^2 \rangle \)
- 3rd moment \( \equiv \text{kurtosis} \)
Moments of Distribution Functions

Generalizing, the $n^{th}$ moment is

$$x_n \equiv \langle x^n \rangle = n\text{th moment} = \frac{\int_{-\infty}^{+\infty} x^n g(x) dx}{\int_{-\infty}^{+\infty} g(x) dx} = \int_{-\infty}^{+\infty} x^n f(x) dx$$

$O^{th}$ moment $\equiv N$

$1^{st}$ moment $\equiv \bar{x}$

$2^{nd}$ moment $\equiv \langle (x - \bar{x})^2 \rangle \rightarrow \langle x^2 \rangle$

$3^{rd}$ moment $\equiv$ kurtosis

The $n^{th}$ moment about the mean is

$$\mu_n \equiv \langle (x - \bar{x})^n \rangle = n\text{th moment about the mean}$$

$$= \frac{\int_{-\infty}^{+\infty} (x-\bar{x})^n g(x) dx}{\int_{-\infty}^{+\infty} g(x) dx} = \int_{-\infty}^{+\infty} (x - \bar{x})^n f(x) dx$$

The standard deviation (or second moment about the mean) is

$$\sigma_x^2 \equiv \mu_2 \equiv \langle (x - \bar{x})^2 \rangle = 2\text{nd moment about the mean}$$

$$= \frac{\int_{-\infty}^{+\infty} (x-\bar{x})^2 g(x) dx}{\int_{-\infty}^{+\infty} g(x) dx} = \int_{-\infty}^{+\infty} (x - \bar{x})^2 f(x) dx$$
Consider a mass on a spring with frequency ω and equilibrium position x₀.

The equations of motion are:

\[ x(t) = A \sin \omega t + x_0 \]
\[ \dot{x}(t) = -A \omega \cos \omega t \]
\[ \ddot{x}(t) = -A \omega^2 \sin \omega t \]

Example of Continuous Distribution Functions and Expectation Values

Harmonic Oscillator: Example from Mechanics

The equations of motion are:

The equations of motion are:

\[ x(t) = A \sin \omega t + x_0 \]
\[ \dot{x}(t) = -A \omega \cos \omega t \]
\[ \ddot{x}(t) = -A \omega^2 \sin \omega t \]

Expected Values

The expectation value of a function \( R(x) \) is

\[
\langle R(x) \rangle \equiv \frac{\int_{-\infty}^{+\infty} R(x) g(x) dx}{\int_{-\infty}^{+\infty} g(x) dx}
\]

\[ = \int_{-\infty}^{+\infty} R(x) f(x) dx \]
Example of Continuous Distribution Functions and Expectation Values

Boltzmann Distribution: Example from Kinetic Theory

Expected Values

The expectation value of a function \( R(x) \) is

\[
\langle R(x) \rangle \equiv \frac{\int_{-\infty}^{+\infty} R(x)g(x)dx}{\int_{-\infty}^{+\infty} g(x)dx} = \int_{-\infty}^{+\infty} R(x)f(x)dx
\]

The Boltzmann distribution function for velocities of particles as a function of temperature, T is:

\[
P(v; T) = 4\pi \left(\frac{M}{2\pi k_B T}\right)^{3/2} v^2 \exp \left[\frac{1}{2} \frac{Mv^2}{k_B T}\right]
\]

Then

\[
\langle v \rangle = \int_{-\infty}^{+\infty} v P(v) dv = \left[8 \frac{k_B T}{\pi M}\right]^{1/2}
\]

\[
\langle v^2 \rangle = \int_{-\infty}^{+\infty} v^2 P(v) dv = \left[3 \frac{k_B T}{M}\right]^{1/2}
\]

implies \( \langle \text{KE} \rangle = \frac{1}{2} M \langle v^2 \rangle = \frac{3}{2} k_B T \)

\[
v_{\text{peak}} = \sqrt{\left[2 \frac{k_B T}{M}\right]^{1/2}} = \left[\frac{2}{3}\right]^{1/2} \langle v^2 \rangle
\]
Example of Continuous Distribution Functions and Expectation Values

Fermi-Dirac Distribution: Example from Kinetic Theory

For a system of identical fermions, the average number of fermions in a single-particle state \( i \), is given by the Fermi–Dirac (F–D) distribution,

\[
\bar{n}_i = \frac{1}{e^{(\epsilon_i - \mu)/kT} + 1}
\]

where \( k_B \) is Boltzmann's constant, \( T \) is the absolute temperature, \( \epsilon_i \) is the energy of the single-particle state \( i \), and \( \mu \) is the total chemical potential.

Since the F–D distribution was derived using the Pauli exclusion principle, which allows at most one electron to occupy each possible state, a result is that \( 0 < \bar{n}_i < 1 \)

When a quasi-continuum of energies \( \epsilon \) has an associated density of states \( g(\epsilon) \) (i.e. the number of states per unit energy range per unit volume) the average number of fermions per unit energy range per unit volume is,

\[
\bar{N}(\epsilon) = g(\epsilon) F(\epsilon)
\]

where \( F(\epsilon) \) is called the Fermi function

\[
F(\epsilon) = \frac{1}{e^{(\epsilon - \mu)/kT} + 1}
\]

so that,

\[
\bar{N}(\epsilon) = \frac{g(\epsilon)}{e^{(\epsilon - \mu)/kT} + 1}
\]
Example of Continuous Distribution Functions and Expectation Values

Finite Square Well: Example from Quantum Mechanics

**Expectation Values**

The expectation value of a QM operator $O(x)$ is

$$\langle O(x) \rangle \equiv \frac{\int_{-\infty}^{+\infty} \Psi^*(x) O(x) \Psi(x) dx}{\int_{-\infty}^{+\infty} \Psi^*(x) \Psi(x) dx}$$

For a finite square well of width $L$, $\Psi_n(x) = \sqrt{2/L} \sin \left[ \frac{n \pi x}{L} \right]$,

$$\langle \Psi_n^*(x) | \Psi_n(x) \rangle \equiv \frac{\int_{-\infty}^{+\infty} \Psi_n^*(x) O(x) \Psi_n(x) dx}{\int_{-\infty}^{+\infty} \Psi_n^*(x) \Psi_n(x) dx} = 1$$

$$\langle x \rangle = \langle \Psi_n^*(x) | x | \Psi_n(x) \rangle \equiv \frac{\int_{-\infty}^{+\infty} \Psi_n^*(x) x \ Psi_n(x) dx}{\int_{-\infty}^{+\infty} \Psi_n^*(x) \Psi_n(x) dx} = L/2$$

$$\langle p \rangle = \langle \Psi_n^*(x) | \frac{\hbar}{i} \frac{\partial}{\partial x} | \Psi_n(x) \rangle \equiv \frac{\int_{-\infty}^{+\infty} \Psi_n^*(x) \frac{\hbar}{i} \frac{\partial}{\partial x} \Psi_n(x) dx}{\int_{-\infty}^{+\infty} \Psi_n^*(x) \Psi_n(x) dx} = 0$$

$$\langle E_n \rangle = \langle \Psi_n^*(x) | i\hbar \frac{\partial}{\partial t} | \Psi_n(x) \rangle \equiv \frac{\int_{-\infty}^{+\infty} \Psi_n^*(x) i\hbar \frac{\partial}{\partial t} \Psi_n(x) dx}{\int_{-\infty}^{+\infty} \Psi_n^*(x) \Psi_n(x) dx} = \frac{n^2 \pi^2 \hbar^2}{2 m L^2}$$
Summary of Distribution Functions

**Introduction**

**Probability Function (Discrete Case)**

The random variable $X$ will be called a discrete random variable if there exists a function $f$ such that $f(x) \geq 0$ and $\sum f(x) = 1$ for $i = 1, 2, 3, \ldots$ and such that for any event $E$,

$$ P(E) = P(X \text{ is in } E) = \sum f(x) $$

where $\sum f(x)$ means sum $f(x)$ over those values $x_i$ that are in $E$ and where $f(x) = P(X = x)$.

The probability that the value of $X$ is some real number $x$, is given by $f(x) = P(X = x)$, where $f$ is called the probability function of the random variable $X$.

**Cumulative Distribution Function (Discrete Case)**

The probability that the value of a random variable $X$ is less than or equal to some real number $x$ is defined as

$$ P(x) = P(X \leq x) = \sum f(x_i), \quad -\infty < x < \infty, $$

where the summation extends over those values of $i$ such that $x_i \leq x$.

**Probability Density (Continuous Case)**

The random variable $X$ will be called a continuous random variable if there exists a function $f$ such that $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x) \, dx = 1$ for all $x$ in interval $-\infty < x < \infty$ and such that for any event $E$

$$ P(E) = P(X \text{ is in } E) = \int_E f(x) \, dx. $$

$f(x)$ is called the probability density of the random variable $X$. The probability that $X$ assumes any given value of $x$ is equal to zero and the probability that it assumes a value on the interval from $a$ to $b$, including or excluding either end point, is equal to

$$ \int_a^b f(x) \, dx. $$

**Cumulative Distribution Function (Continuous Case)**

The probability that the value of a random variable $X$ is less than or equal to some real number $x$ is defined as

$$ F(x) = P(X \leq x), \quad -\infty < x < \infty $$

From the cumulative distribution, the density, if it exists, can be found from

$$ f(x) = \frac{dF(x)}{dx}. $$

From the cumulative distribution

$$ P(a \leq X \leq b) = P(X \leq b) - P(X \leq a) = P(b) - P(a) $$

**Mathematical Expectation**

**A. Expected Value**

Let $X$ be a random variable with probability density $f(x)$. Then the expected value of $X$, $E(X)$, is defined to be

$$ E(X) = \sum xf(x) $$

if $X$ is discrete and

$$ E(X) = \int_{-\infty}^{\infty} xf(x) \, dx $$

if $X$ is continuous. The expected value of a function $g$ of a random variable $X$ is defined as

$$ E[g(X)] = \sum g(x) \cdot f(x) $$

if $X$ is discrete and

$$ E[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f(x) \, dx $$

if $X$ is continuous.

**Theorems**

1. $E(AX + By) = aE(X) + bE(Y)$
2. $E(X \cdot Y) = E(X) \cdot E(Y)$ if $X$ and $Y$ are statistically independent.

**B. Moments**

- a. Moments About the Origin. The moments about the origin of a probability distribution are the expected values of the random variable which has the given distribution. The $r$th moment of $X$, usually denoted by $\mu_r$, is defined as

$$ \mu_r = E[X^r] = \sum x^r f(x) $$

if $X$ is discrete and

$$ \mu_r = E[X^r] = \int_{-\infty}^{\infty} x^r f(x) \, dx $$

if $X$ is continuous.

The first moment, $\mu_1$, is called the mean of the random variable $X$ and is usually denoted by $\mu$.

- b. Moments About the Mean. The $r$th moment about the mean, usually denoted by $\mu_r$, is defined as

$$ \mu_r = E[(X - \mu)^r] = \sum (x - \mu)^r f(x) $$

if $X$ is discrete and

$$ \mu_r = E[(X - \mu)^r] = \int_{-\infty}^{\infty} (x - \mu)^r f(x) \, dx $$

if $X$ is continuous.

The second moment about the mean, $\mu_2$, is given by

$$ \mu_2 = E[(X - \mu)^2] = \mu_2 - \mu^2 $$

and is called the variance of the random variable $X$, and is denoted by $\sigma^2$. The square root of the variance, $\sigma$, is called the standard deviation.

**Theorems**

1. $\sigma^2 = \mu_2 - \mu^2$
2. $\mu_{r+2} = \mu_r + \mu_{r+2}^2$
3. $\sigma_{r+2} = \sigma_r + \sigma_{r+2}$
The Gaussian Distribution Function

References: Taylor Ch. 5
Gaussian Integrals

Factorial Approximations

\[ n! \approx (2\pi n)^{1/2} n^n \exp \left[ -n + \frac{1}{12n} + O \left( \frac{1}{n^2} \right) \right] \]

\[ \log(n!) \approx \frac{1}{2} \log(2\pi) + \left( n + \frac{1}{2} \right) \log(n) - n + \frac{1}{12n} + O \left( \frac{1}{n^2} \right) \]

\[ \log(n!) \approx n \log(n) - n \quad \text{(for all terms decreasing faster than linearly with n)} \]

Gaussian Integrals

\[ I_m = 2 \int_0^\infty x^m \exp[-x^2] \, dx \quad ; \, m > -1 \]

\[ I_m = 2 \int_0^\infty y^n \exp[-y] \, dy \equiv \Gamma(n + 1) \quad ; \, x^2 \equiv y, \, 2 \, dx = y^{1/2} \, dy, \quad n \equiv \frac{1}{2} (m - 1) \]

\[ I_0 = \Gamma \left( n = \frac{1}{2} \right) = \sqrt{\pi} \quad ; \, m=0, \quad n = -\frac{1}{2} \]

\[ I_{2k} = \Gamma \left( k + \frac{1}{2} \right) = (k - 1/2)(k - 3/2)\ldots\left(3/2\right)(1/2) \sqrt{\pi} \quad ; \, \text{even } m \quad m=2k>0, \quad n = k - \frac{1}{2} \]

\[ I_{2k+1} = \Gamma(k + 1) = k! \quad ; \, \text{odd } m \quad m=2k+1>0, \quad n = k \geq 0 \]
Gaussian Distribution Function

\[ G_{X,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-X)^2}{2\sigma^2}} \]

- Distribution Function
- Independent Variable
- Center of Distribution (mean)
- Width of Distribution (standard deviation)
- Normalization Constant

Figure 5.10. Two normal, or Gauss, distributions.
Effects of Increasing N on Gaussian Distribution Functions

Consider the limiting distribution function as $N \to \infty$ and $dx \to 0$
### Defining the Gaussian distribution function in Mathcad

**Measure Data:**

Normal Distribution Parameters:
- **Mean:** \( \mu = 10.000 \)
- **Standard Deviation:** \( \sigma = 5.000 \)

Number of "Measurements":
- \( N = 6 \)

Measurements:
- \( x := \text{norn}(N, \mu, \sigma) \)

**Histogram Calculations:**

- Number of intervals: \( M := \text{floor}(N^{2.5}) + 2 \)
- M = 4.000
- \( m := 0 \ldots (M) \)

- Interval spacing: \( \Delta x := \left( \frac{\text{ceil}(\text{max}(x)) - \text{floor}(\text{min}(x))}{M} \right) \)
- \( \Delta x = 2.500 \)

Calculate Intervals:
- \( \text{Int}_m := \text{floor}(\text{min}(x)) + m \cdot \Delta x \)

Calculate Frequencies:
- \( F := \text{hist}(\text{Int}, x) \)

**Gaussian Distribution Function:**

Define distribution function:
- \( \text{Norm}(A, \mu, \sigma, x) := \frac{A}{\sigma \sqrt{2 \cdot \pi}} \cdot \exp \left[ \frac{-(x - \mu)^2}{2 \cdot \sigma^2} \right] \)

Maximum of distribution function:
- \( N_{\text{max}} := \text{Norm}(1, \mu, \sigma, x) \)
- \( N_{\text{max}} = 0.080 \)

I suggest you “investigate” these with the Mathcad sheet on the web site.
Using Mathcad to define other common distribution functions.

Consider the limiting distribution function as $N \to \infty$ and $\Delta_k \to 0$. 

**Alternative Distribution Function:**

<table>
<thead>
<tr>
<th>Distribution function</th>
<th>Normal Distribution Parameters</th>
<th>&quot;Measurements&quot;</th>
<th>Frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binomial</td>
<td>$p := 0.68$ $\nu := 15$</td>
<td>$x_b := \text{rbinom}(N, \nu, p)$</td>
<td>$F_b := \text{hist}($Int$\cdot x_b)$</td>
</tr>
<tr>
<td>Poisson</td>
<td>$\lambda := 10$</td>
<td>$x_p := \text{rpois}(N, \lambda)$</td>
<td>$F_p := \text{hist}($Int$\cdot x_p)$</td>
</tr>
<tr>
<td>Cauchy</td>
<td>$\mu := \mu$ $\sigma := \sigma$</td>
<td>$x_c := \text{rcauchy}(N, 1, s)$</td>
<td>$F_c := \text{hist}($Int$\cdot x_c)$</td>
</tr>
<tr>
<td>Chi-squared</td>
<td>$d := N - 4$</td>
<td>$x_\chi := \text{rpois}(N, \lambda)$</td>
<td>$F_\chi := \text{hist}($Int$\cdot x_\chi)$</td>
</tr>
</tbody>
</table>
Consider the Gaussian distribution function

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

Use the normalization condition to evaluate the normalization constant (see Taylor, p. 132)

The mean, \( \bar{X} \), is the first moment of the Gaussian distribution function (see Taylor, p. 134)

The standard deviation, \( \sigma_x \), is the standard deviation of the mean of the Gaussian distribution function (see Taylor, p. 143)
When is mean $x$ not $X_{\text{best}}$?

Answer: When the distribution is not symmetric about $X$.

Example: Cauchy Distribution
When is mean x not $X_{\text{best}}$?

Maxwell speed distribution law is

$$P(v) = 4\pi \left( \frac{M}{2\pi RT} \right)^{3/2} v^2 e^{-Mv^2/2RT}.$$  

There are three candidates for what is called the "average" value of the speed of the Maxwell speed distribution.

Firstly, by finding the maximum of the MSD (by differentiating, setting the derivative equal to zero and solving for the speed), we can determine the most probable speed. Calling this $v_{mp}$, we find that:

$$v_{mp} = \left( \frac{2kT}{m} \right)^{1/2}.$$

Second, we can find the mean value of v from the MSD. Calling this $\bar{v}$:

$$\bar{v} = \left( \frac{8kT}{\pi m} \right)^{1/2}.$$

Third and finally, we can find the root mean square of the speed by finding the expected value of $v^2$. (Alternatively, and much simpler, we can solve it by using the equipartition theorem.) Calling this $v_{rms}$:

$$v_{rms} = \left( \frac{3kT}{m} \right)^{1/2}.$$

Notice that $v_{mp} < \bar{v} < v_{rms}$.

These are three different ways of defining the average velocity, and they are not numerically the same. It is important to be clear about which quantity is being used.
When is mean $x$ not $X_{\text{best}}$?

Answer: When the distribution is has more than one peak.
The Gaussian Distribution Function and Its Relation to Errors
A Review of Probabilities in Combination

1 head AND 1 Four
\[ P(H, 4) = P(H) \times P(4) \]

1 Six OR 1 Four
\[ P(6, 4) = P(6) + P(4) \]
(true for a “mutually exclusive” single role)

1 head OR 1 Four
\[ P(H, 4) = P(H) + P(4) - P(H \text{ and } 4) \]
(true for a “non-mutually exclusive” events)

NOT 1 Six
\[ P(\text{NOT 6}) = 1 - P(6) \]

Probability of a data set of \( N \) like measurements, \((x_1, x_2, \ldots, x_N)\)
\[ P(x_1, x_2, \ldots, x_N) = P(x_1) \times P(x_2) \times \ldots P(x_N) \]
We will use the Gaussian distribution as applied to random variables to develop the mathematical basis for:

- Mean
- Standard deviation
- Standard deviation of the mean (SDOM)
- Moments and expectation values
- Error propagation formulas
- Addition of errors in quadrature (for independent and random measurements)
- Numerical values for confidence limits (t-test)
- Principle of maximal likelihood
- Central limit theorem
- Weighted distributions and Chi squared
- Schwartz inequality (i.e., the uncertainty principle) (next lecture)
Consider the Gaussian distribution function

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

Use the normalization condition to evaluate the normalization constant (see Taylor, p. 132)

\[
1 = \int_{-\infty}^{\infty} G_{\bar{X}\sigma}(x) \, dx = \int_{-\infty}^{\infty} N \exp\left[-\frac{(x - \bar{X})^2}{2\sigma^2}\right] \, dx
\]

\[
1 = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{y^2}{2\sigma^2}\right] \, dy
\]

\[
N = 1/(\sigma\sqrt{2\pi})
\]

The mean, \(\bar{X}\), is the first moment of the Gaussian distribution function (see Taylor, p. 134)

\[
\langle x \rangle = \int_{-\infty}^{\infty} x \, G_{\bar{X}\sigma}(x) \, dx = \bar{X}
\]

The standard deviation, \(\sigma_x\), is the standard deviation of the mean of the Gaussian distribution function (see Taylor, p. 143)

\[
\sigma_x^2 = \int_{-\infty}^{\infty} (x - \bar{X})^2 G_{\bar{X}\sigma}(x) \, dx = \sigma^2
\]
Standard Deviation of Gaussian Distribution

\[ \text{Prob(within } \sigma) = \int_{x-\sigma}^{x+\sigma} G_{x,\sigma}(x) \, dx \]

\[ = \frac{1}{\sigma \sqrt{2\pi}} \int_{x-\sigma}^{x+\sigma} \exp\left(-\frac{(x-x)^2}{2\sigma^2}\right) \, dx. \]

\[ \text{Area under curve (probability that } -\sigma < x < +\sigma \text{) is 68%.} \]

Ah, that’s highly significant!

1% or ~3\( \sigma \) “Highly Significant”
5% or ~2\( \sigma \) “Significant”
1\( \sigma \) “Within errors”

5 ppm or ~5\( \sigma \) “Valid for HEP”

See Sec. 10.6: Testing of Hypotheses

Figure 5.13. The probability \( \text{Prob(} \text{within } \sigma \text{)} \) that a measurement of \( x \) will fall within \( \sigma \) standard deviations of the true value \( x = X \). Two common names for this function are the normal error integral and the error function, \( \text{erf} \).

More complete Table in App. A and B
Error Function of Gaussian Distribution

Error Function: (probability that $-t\sigma < x < +t\sigma$).

$$Prob(wi\text{thin } t\sigma) = \frac{1}{\sqrt{2\pi}} \int_{-t}^{t} e^{-z^2/2} \, dz.$$  \hspace{1cm} (5.35)

Area under curve (probability that $-t\sigma < x < +t\sigma$) is given in Table at right.

Complementary Error Function: (probability that $-x < -t\sigma \text{ AND } x < +t\sigma$).

$$Prob(x \text{ outside } t\sigma) = 1 - Prob(x \text{ within } t\sigma)$$

Probable Error: (probability that $-0.67\sigma < x < +0.67\sigma$) is 50%.

Error Function: Tabulated values-see App. A.

Probable Error: (probability that $-0.67\sigma < x < +0.67\sigma$) is 50%.

More complete Table in App. A and B

---

**Figure 5.13.** The probability $Prob(wi\text{thin } t\sigma)$ that a measurement of $x$ will fall within $t$ standard deviations of the true value $x = X$. Two common names for this function are the normal error integral and the error function, erf(t).
Useful Points on Gaussian Distribution

**Full Width at Half Maximum**
FWHM
(See Prob. 5.12)

\[
\text{FWHM} = 2\sigma \sqrt{2 \ln 2} = 2.35 \sigma.
\]

**Points of Inflection**
Occur at \( \pm \sigma \)
(See Prob. 5.13)

---

**Figure 5.20.** The points \( X \pm \sigma \) are the points of inflection of the Gauss curve; for Problem 5.13.
Error Analysis and Gaussian Distribution

Adding a Constant

\[ q = x + A, \quad \text{(5.47)} \]

\[ G_{x,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-x)^2}{2\sigma^2}} \]

(probability of obtaining value \( q \)) \( \propto e^{-\frac{(q-A-x)^2}{2\sigma^2}} \]

\[ = e^{-\frac{(q-(x+A))^2}{2\sigma^2}}. \quad (5.49) \]

Multiplying by a Constant

\[ q = Bx, \]

\[ G_{x,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-x)^2}{2\sigma^2}} \]

(probability of obtaining value \( q \)) \( \propto \exp\left[-\left(\frac{q}{B} - x\right)^2/2\sigma_x^2\right] \]

\[ = \exp\left[-\left(q - BX\right)^2/2B^2\sigma_x^2\right]. \quad (5.50) \]
Error Propagation: Addition

Sum of Two Variables

Consider the derived quantity
\( Z = X + Y \)
(with \( X = 0 \) and \( Y = 0 \))

\[
\begin{align*}
\text{Prob}(x) & \propto \exp\left(\frac{-x^2}{2\sigma_x^2}\right) \\
\text{Prob}(y) & \propto \exp\left(\frac{-y^2}{2\sigma_y^2}\right).
\end{align*}
\]

Error in \( Z \):
Multiple two probabilities

\[
\text{Prob}(x, y) = \text{Prob}(x) \cdot \text{Prob}(x) \propto \exp\left(-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)\right)
\]

\[
\propto \exp\left(-\frac{1}{2}\left(\frac{(x+y)^2}{\sigma_x^2 + \sigma_y^2}\right)\right)
\]

\[
\propto \exp\left(-\frac{1}{2}\left(\frac{(x+y)^2}{\sigma_x^2 + \sigma_y^2}\right) + \frac{z^2}{2}\right)
\]

\( X + Y \rightarrow Z \) and \( \sigma_x^2 + \sigma_y^2 \rightarrow \sigma_z^2 \)
(addition in quadrature for random, independent variables!)

ICBST (Eq. 5.53)

Integrates to \( \sqrt{2\pi} \)
General Formula for Error Propagation

How do we determine the error of a derived quantity \( Z(X,Y,...) \) from errors in \( X,Y,... \)?

General formula for error propagation  see [Taylor, Secs. 3.5 and 3.9]

Uncertainty as a function of one variable  [Taylor, Sec. 3.5]

1. Consider a graphical method of estimating error
   a) Consider an arbitrary function \( q(x) \)
   b) Plot \( q(x) \) vs. \( x \).
   c) On the graph, label:
      (1) \( q_{\text{best}} = q(x_{\text{best}}) \)
      (2) \( q_{\text{hi}} = q(x_{\text{best}} + \delta x) \)
      (3) \( q_{\text{low}} = q(x_{\text{best}} - \delta x) \)
   d) Making a linear approximation:
      \[
      q_{\text{hi}} = q_{\text{best}} + \text{slope} \cdot \delta x = q_{\text{best}} + \left( \frac{\partial q}{\partial x} \right) \delta x \\
      q_{\text{low}} = q_{\text{best}} - \text{slope} \cdot \delta x = q_{\text{best}} - \left( \frac{\partial q}{\partial x} \right) \delta x
      \]
   e) Therefore:
      \[
      \delta q = \left| \frac{\partial q}{\partial x} \right| \cdot \delta x
      \]
      Note the absolute value.
General formula for uncertainty of a function of one variable

\[
\delta q = \left| \frac{\partial q}{\partial x} \right| \delta x \quad \text{[Taylor, Eq. 3.23]}
\]

Can you now derive for specific rules of error propagation:

1. Addition and Subtraction \hspace{1cm} \text{[Taylor, p. 49]}
2. Multiplication and Division \hspace{1cm} \text{[Taylor, p. 51]}
3. Multiplication by a constant (exact number) \hspace{1cm} \text{[Taylor, p. 54]}
4. Exponentiation (powers) \hspace{1cm} \text{[Taylor, p. 56]}
General Formula for Multiple Variables

Uncertainty of a function of multiple variables [Taylor, Sec. 3.11]

1. It can easily (no, really) be shown that (see Taylor Sec. 3.11) for a function of several variables

\[ \delta q(x, y, z, ...) = \left| \frac{\partial q}{\partial x} \right| \delta x + \left| \frac{\partial q}{\partial y} \right| \delta y + \left| \frac{\partial q}{\partial z} \right| \delta z + \ldots \]  
   [Taylor, Eq. 3.47]

2. More correctly, it can be shown that (see Taylor Sec. 3.11) for a function of several variables

\[ \delta q(x, y, z, ...) \leq \left| \frac{\partial q}{\partial x} \right| \delta x + \left| \frac{\partial q}{\partial y} \right| \delta y + \left| \frac{\partial q}{\partial z} \right| \delta z + \ldots \]  
   [Taylor, Eq. 3.47]

where the equals sign represents an upper bound, as discussed above.

3. For a function of several independent and random variables

\[ \delta q(x, y, z, ...) = \sqrt{ \left( \frac{\partial q}{\partial x} \cdot \delta x \right)^2 + \left( \frac{\partial q}{\partial y} \cdot \delta y \right)^2 + \left( \frac{\partial q}{\partial z} \cdot \delta z \right)^2 + \ldots } \]  
   [Taylor, Eq. 3.48]

Again, the proof is left for Ch. 5.
Error Propagation: General Case

How do we determine the error of a derived quantity \( Z(X,Y,...) \) from errors in \( X,Y,... \)?

Consider the arbitrary derived quantity \( q(x,y) \) of two independent random variables \( x \) and \( y \).

Expand \( q(x,y) \) in a Taylor series about the expected values of \( x \) and \( y \) (i.e., at points near \( X \) and \( Y \)).

\[
q(x, y) = q(X, Y) + \left( \frac{\partial q}{\partial x} \right)_X (x - X) + \left( \frac{\partial q}{\partial y} \right)_Y (y - Y)
\]

Product of Two Variables

\[
\delta q(x, y) = \sigma_q = \sqrt{q(X, Y) + \left[ \left( \frac{\partial q}{\partial x} \right)_X \sigma_x \right]^2 + \left[ \left( \frac{\partial q}{\partial y} \right)_Y \sigma_y \right]^2}
\]
**SDOM of Gaussian Distribution**

**Standard Deviation of the Mean**

Each measurement has similar $\sigma_{x_i} = \sigma_{\bar{x}}$

and similar partial derivatives

Thus...

The SDOM decreases as the square root of the number of measurements.

That is, the relative width, $\sigma/\bar{x}$, of the distribution gets narrower as more measurements are made.
Two Key Theorems from Probability

Central Limit Theorem
For random, independent measurements (each with a well-defined expectation value and well-defined variance), the arithmetic mean (average) will be approximately normally distributed.

Principle of Maximum Likelihood
Given the N observed measurements, \( x_1, x_2, \ldots x_N \), the best estimates for \( \bar{X} \) and \( \sigma \) are those values for which the observed \( x_1, x_2, \ldots x_N \) are most likely.
Mean of Gaussian Distribution as “Best Estimate”

Principle of Maximum Likelihood

To find the most likely value of the mean (the best estimate of \( \dot{x} \)), find \( X \) that yields the highest probability for the data set.

Consider a data set \( \{x_1, x_2, x_3 \ldots x_N \} \)

Each randomly distributed with

\[
Prob_{X,\sigma}(x_i) = G_{X,\sigma}(x_i) \equiv \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x_i - X)^2}{2\sigma}} \propto \frac{1}{\sigma} e^{-\frac{(x_i - X)^2}{2\sigma}}
\]

The combined probability for the full data set is the product

\[
Prob_{X,\sigma}(x_1, x_2 \ldots x_N) = Prob_{X,\sigma}(x_1) \times Prob_{X,\sigma}(x_2) \times \ldots \times Prob_{X,\sigma}(x_N)
\]

\[
\propto \frac{1}{\sigma} e^{-\frac{(x_1 - X)^2}{2\sigma}} \times \frac{1}{\sigma} e^{-\frac{(x_2 - X)^2}{2\sigma}} \times \ldots \times \frac{1}{\sigma} e^{-\frac{(x_N - X)^2}{2\sigma}} = \frac{1}{\sigma^N} e^{\sum_{i=1}^{N} (x_i - X)^2 / 2\sigma}
\]

Best Estimate of \( X \) is from maximum probability or minimum summation

**Minimize**

\[
\sum_{i=1}^{N} \frac{(x_i - X)^2}{\sigma}
\]

**Solve for derivative wrst**

\[X \text{ set to 0} \]

\[\sum_{i=1}^{N} (x_i - X) = 0 \]

**Best estimate of \( X \)**

\[X_{\text{best}} = \frac{\sum x_i}{N}\]
Uncertainty of “Best Estimates” of Gaussian Distribution

**Principle of Maximum Likelihood**

To find the most likely value of the standard deviation (the best estimate of the width of the x distribution), find $\sigma_x$ that yields the highest probability for the data set.

Consider a data set $\{x_1, x_2, x_3 \ldots x_N\}$

The combined probability for the full data set is the product

$$\text{Prob}_{x,\sigma}(x_1, x_2\ldots x_N) = \text{Prob}_{x,\sigma}(x_1) \times \text{Prob}_{x,\sigma}(x_2) \times \ldots \times \text{Prob}_{x,\sigma}(x_N)$$

$$\propto \frac{1}{\sigma} e^{-\frac{(x_1-X)^2}{2\sigma}} \times \frac{1}{\sigma} e^{-\frac{(x_2-X)^2}{2\sigma}} \times \ldots \times \frac{1}{\sigma} e^{-\frac{(x_N-X)^2}{2\sigma}} = \frac{1}{\sigma^N} e^{\sum-(x_i-X)^2/2\sigma}$$

**Best Estimate of X is from maximum probability or minimum summation**

Minimize

$$\sum_{i=1}^{N} \frac{(x_i - X)^2}{\sigma}$$

Solve for derivative wrst X set to 0

$$\sum_{i=1}^{N} (x_i - X) = 0$$

$X_{\text{best}} = \frac{\sum x_i}{N}$

**Best Estimate of $\sigma$ is from maximum probability or minimum summation**

Minimize

$$\sum_{i=1}^{N} \frac{(x_i - X)^2}{\sigma}$$

Solve for derivative wrst $\sigma$ set to 0

See Prob. 5.26

$\sigma_{\text{best}} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \frac{(x_i - X)^2}{\sigma}}$
Combining Data Sets

Weighted Averages

References: Taylor Ch. 7
Question: How can we properly combine two or more separate independent measurements of the same randomly distributed quantity to determine a best combined value with uncertainty?
Consider two measurements of the same quantity, described by a random Gaussian distribution:

\[ <x_1> \pm \sigma_{x1} \quad \text{and} \quad <x_2> \pm \sigma_{x2} \]

Assume negligible systematic errors.

The probability of measuring two such measurements is

\[
Prob_x(x_1, x_2) = Prob_x(x_1) \cdot Prob_x(x_2) = \frac{1}{\sigma_1 \sigma_2} e^{-\chi^2/2} \quad \text{where} \quad \chi^2 \equiv \left[ \frac{(x_1 - X)}{\sigma_1} \right]^2 + \left[ \frac{(x_2 - X)}{\sigma_2} \right]^2
\]

To find the best value for \( \chi \), find the maximum Prob or minimum \( \chi^2 \)

Note: \( \chi^2 \), or Chi squared, is the sum of the squares of the deviations from the mean, divided by the corresponding uncertainty.

Such methods are called “Methods of Least Squares”. They follow directly from the Principle of Maximum Likelihood.
Weighted Averages

The probability of measuring two such measurements is

\[ \text{Prob}_x(x_1, x_2) = \text{Prob}_x(x_1) \text{ Prob}_x(x_2) \]

\[ = \frac{1}{\sigma_1 \sigma_2} e^{-\chi^2/2} \quad \text{where} \quad \chi^2 \equiv \left[ \frac{(x_1 - X)}{\sigma_1} \right]^2 + \left[ \frac{(x_2 - X)}{\sigma_2} \right]^2 \]

To find the best value for \( \chi \), find the maximum Prob or minimum \( \chi^2 \)

**Best Estimate of \( \chi \) is from maximum probibility or minimum summation**

Minimize Sum | Solve for derivative wrst \( \chi \) set to 0 | Solve for best estimate of \( \chi \)
---|---|---
\[ \chi^2 \equiv \left[ \frac{(x_1 - X)}{\sigma_1} \right]^2 + \left[ \frac{(x_2 - X)}{\sigma_2} \right]^2 \]
\[ 2 \left[ \frac{(x_1 - X)}{\sigma_1} \right] + 2 \left[ \frac{(x_2 - X)}{\sigma_2} \right] = 0 \]
\[ X_{\text{best}} = \left( \frac{x_1}{\sigma_1^2 + \sigma_2^2} \right) \left( \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right) \]

This leads to

\[ X_{\text{W-avg}} = \frac{w_1 x_1 + w_2 x_2}{w_1 + w_2} = \frac{\sum w_i x_i}{\sum w_i} \quad \text{where} \quad w_i = \frac{1}{(\sigma_i)^2} \]

**Note:** If \( w_1 = w_2 \), we recover the standard result \( X_{\text{wavg}} = (1/2) (x_1 + x_2) \)

Finally, the width of a weighted average distribution is

\[ \sigma_{\text{weighted avg}} = \frac{1}{\sum_i w_i} \]
**Weighted Averages on Steroids**

A very powerful method for combining data from different sources with different methods and uncertainties (or, indeed, data of related measured and calculated quantities) is Kalman filtering.

The Kalman filter, also known as linear quadratic estimation (LQE), is an algorithm that uses a series of measurements observed over time, containing noise (random variations) and other inaccuracies, and produces estimates of unknown variables that tend to be more precise than those based on a single measurement alone.

The Kalman filter keeps track of the estimated state of the system and the variance or uncertainty of the estimate. The estimate is updated using a state transition model and measurements. $x_{k|k-1}$ denotes the estimate of the system's state at time step $k$ before the $k^{th}$ measurement $y_k$ has been taken into account; $P_{k|k-1}$ is the corresponding uncertainty. --Wikipedia, 2013.

Ludger Scherliess, of USU Physics, is a world expert at using Kalman filtering for the assimilation of satellite and ground-based data and the USU GAMES model to predict space weather.
Rejecting Data

Chauvenet’s Criterion

References: Taylor Ch. 6
What is a good criteria for rejecting data?

Question: When is it “reasonable” to discard a seemingly “unreasonable” data point from a set of randomly distributed measurements?

- Never
- Whenever it makes things look better
- Chauvenet’s criterion provides a (quantitative) compromise
Question: When is it “reasonable” to discard a seemingly “unreasonable” data point from a set of randomly distributed measurements?

Often in physics, experimental observations are termed “anomalous” before they are understood. Once theory succeeds in explaining and illuminating the observations, they are no longer “anomalous” and instead come to be regarded as “obvious.” A crucial paper can trigger such an “anomalous \rightarrow obvious” transition, and in the present case that key role was played by a 1975 paper by Scher and Montroll. That landmark paper has become basic to our understanding of a striking characteristic of carrier motion (now called dispersive transport) which is a common occurrence in amorphous semiconductors, though foreign to our experience with crystals.
Question: When is it “reasonable” to discard a seemingly “unreasonable” data point from a set of randomly distributed measurements?

- Whenever it makes things look better

**Disney’s First Law**

Wishing will make it so.

**Disney’s Second Law**

Dreams are more colorful than reality.
Rejection of Data

**Chauvenet’s Criterion**

Data may be rejected if the expected number of measurements at least as deviant as the suspect measurement is less than 50%.

Consider a set of $N$ measurements of a single quantity

\[ \{ x_1, x_2, \ldots, x_N \} \]

Calculate $\langle x \rangle$ and $\sigma_x$ and then determine the fractional deviations from the mean of all the points:

\[ x_{frac\_dev} = \frac{|x_i - \bar{x}|}{\sigma_x} \]

For the suspect point(s), $x_{suspect}$, find the probability of such a point occurring in $N$ measurements

\[ n = (\text{expected number as deviant as } x_{suspect}) \]

\[ = N \text{ Prob(outside } x_{suspect} \cdot \sigma_x) \]
Error Function of Gaussian Distribution

Error Function: (probability that $-t\sigma < x < +t\sigma$).

\[ \text{Prob}(\text{within } t\sigma) = \frac{1}{\sqrt{2\pi}} \int_{-t}^{t} e^{-z^2/2} \, dz. \]

(5.35)

Area under curve (probability that $-t\sigma < x < +t\sigma$) is given in Table at right.

Probable Error: (probability that $-0.67\sigma < x < +0.67\sigma$) is 50%.

Figure 5.13. The probability $\text{Prob}(\text{within } t\sigma)$ that a measurement of $x$ will fall within $t$ standard deviations of the true value $x = X$. Two common names for this function are the normal error integral and the error function, erf(t).
Chauvenet’s Criterion

The probability that a data point is likely to fall outside a given deviation is:

\[ \text{Prob}(x_{\text{test}}, X, \sigma) := 1 - \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \int_{-|x_{\text{test}}|}^{+|x_{\text{test}}|} \frac{(x-X)^2}{2 \cdot \sigma^2} \, dx \]

<table>
<thead>
<tr>
<th>( x )</th>
<th>( \text{Frac}_\text{Dev} )</th>
<th>( \text{Prob}(x_i, \overline{x}, \sigma_x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.468</td>
<td>0.68</td>
</tr>
<tr>
<td>1</td>
<td>0.281</td>
<td>0.61</td>
</tr>
<tr>
<td>2</td>
<td>0.019</td>
<td>0.507</td>
</tr>
<tr>
<td>3</td>
<td>2.936</td>
<td>1.66 \times 10^{-3}</td>
</tr>
<tr>
<td>4</td>
<td>0.318</td>
<td>0.625</td>
</tr>
<tr>
<td>5</td>
<td>0.655</td>
<td>0.744</td>
</tr>
<tr>
<td>6</td>
<td>0.58</td>
<td>0.719</td>
</tr>
<tr>
<td>7</td>
<td>0.019</td>
<td>0.493</td>
</tr>
<tr>
<td>8</td>
<td>0.393</td>
<td>0.653</td>
</tr>
<tr>
<td>9</td>
<td>0.243</td>
<td>0.596</td>
</tr>
</tbody>
</table>

Including all data points

- \( \overline{x} := \text{mean}(x) = 46.95 \text{ m} \)
- \( \sigma_x := \text{stddev}(x) = 2.673 \text{ m} \)

Excluding the rejected data point

- \( \overline{x}_{\text{CH}} := \text{mean}(X_{\text{CH}}) = 46.078 \text{ m} \)
- \( \sigma_{x_{\text{CH}}} := \text{stddev}(X_{\text{CH}}) = 0.577 \text{ m} \)
Chauvenet’s Criterion—Example 1

Example: Ten Measurements of a Length

A student makes 10 measurements of one length $x$ and gets the results (all in mm)

$$46, 48, 44, 38, 45, 47, 58, 44, 45, 43.$$  

Noticing that the value 58 seems anomalously large, he checks his records but can find no evidence that the result was caused by a mistake. He therefore applies Chauvenet’s criterion. What does he conclude?

Accepting provisionally all 10 measurements, he computes

$$\bar{x} = 45.8 \quad \text{and} \quad \sigma_x = 5.1.$$
Chauvenet’s Details (1)

The difference between the suspect value $x_{sus} = 58$ and the mean $\bar{x} = 45.8$ is 12.2, or 2.4 standard deviations; that is,

$$t_{sus} = \frac{x_{sus} - \bar{x}}{\sigma_x} = \frac{58 - 45.8}{5.1} = 2.4.$$ 

Referring to the table in Appendix A, he sees that the probability that a measurement will differ from $\bar{x}$ by $2.4\sigma_x$ or more is

$$Prob(\text{outside } 2.4\sigma) = 1 - Prob(\text{within } 2.4\sigma) = 1 - 0.984 = 0.016.$$ 

In 10 measurements, he would therefore expect to find only 0.16 of one measurement as deviant as his suspect result. Because 0.16 is less than the number 0.5 set by Chauvenet’s criterion, he should at least consider rejecting the result.

If he decides to reject the suspect 58, then he must recalculate $\bar{x}$ and $\sigma_x$ as

$$\bar{x} = 44.4 \text{ and } \sigma_x = 2.9.$$ 

As you would expect, his mean changes a bit, and his standard deviation drops appreciably.
Consider the following example of the application of Chauvenet's Criterion to determine if a certain datum should be rejected.

A set of $N=10$ measurements of a length are made. The data are assumed to be described by a random Gaussian distribution.

**Enter Data**

Number of data points: $N := 10$

Data indices: $i := 0..(N - 1)$

Enter data set:

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>m</th>
</tr>
</thead>
<tbody>
<tr>
<td>45.7</td>
<td>m</td>
</tr>
<tr>
<td>46.2</td>
<td>m</td>
</tr>
<tr>
<td>46.9</td>
<td>m</td>
</tr>
<tr>
<td>54.8</td>
<td>m</td>
</tr>
<tr>
<td>46.1</td>
<td>m</td>
</tr>
<tr>
<td>45.2</td>
<td>m</td>
</tr>
<tr>
<td>45.4</td>
<td>m</td>
</tr>
<tr>
<td>47.0</td>
<td>m</td>
</tr>
<tr>
<td>45.9</td>
<td>m</td>
</tr>
<tr>
<td>46.3</td>
<td>m</td>
</tr>
</tbody>
</table>

Calculate mean: $x_{\text{mean}} := \text{mean}(x) = 46.95\text{m}$

Calculate standard deviation: $\sigma_x := \text{stdev}(x) = 2.673\text{m}$

Calculate fractional deviation from the mean:

$\text{Frac}_i := \left| \frac{x_i - x_{\text{mean}}}{\sigma_x} \right|$
To apply Chauvenet's criterion, we first sort the data \( x \) in order of ascending values of the fractional deviation from the mean. The probability that a data point is likely to fall outside a given deviation is then calculated. We then determine how many data that should be eliminated based on Chauvenet's criterion.

Sort data in ascending order:

\[
x_{\text{order}} := \text{csort}\left(\frac{x}{m}, \text{Frac}_\text{Dev}, 1\right)
\]

The probability that a data point is likely to fall outside a given deviation is:

\[
\text{Prob}(x_{\text{test}}, X, \sigma) := 1 - \int_{-|x_{\text{test}}|}^{+|x_{\text{test}}|} \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{-\frac{(x-X)^2}{2\sigma^2}} \, dx
\]

Apply Chauvenet's criterion:

\[
\text{Reject}(x, X, \sigma, N) := \text{if}[(N \cdot \text{Prob}(x, X, \sigma)) > 50 \cdot \% , \text{"Keep"}, \text{"Reject"}]
\]

Determine how many data points should be rejected:

\[
N_{\text{reject}} := \sum_{i = 0}^{N-1} \text{if}[(N \cdot \text{Prob}(x, X, \sigma, x_i)) > 50 \cdot \%, 0, 1] = 1
\]

Chauvenet's Criterion—Details (3)

<table>
<thead>
<tr>
<th>( x )</th>
<th>( \text{Frac}_\text{Dev} % )</th>
<th>( \text{Prob}(x_{\text{test}}, X, \sigma, x) \cdot N \cdot \text{Prob}(x_{\text{mean}}, \sigma, x) = \text{Reject}(x, x_{\text{mean}}, \sigma, x, N) = )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>46.2</td>
<td>0.68</td>
</tr>
<tr>
<td>2</td>
<td>46.9</td>
<td>6.105</td>
</tr>
<tr>
<td>3</td>
<td>54.8</td>
<td>6.247</td>
</tr>
<tr>
<td>4</td>
<td>46.1</td>
<td>7.436</td>
</tr>
<tr>
<td>5</td>
<td>45.2</td>
<td>7.19</td>
</tr>
<tr>
<td>6</td>
<td>45.4</td>
<td>4.925</td>
</tr>
<tr>
<td>7</td>
<td>47</td>
<td>6.528</td>
</tr>
<tr>
<td>8</td>
<td>45.9</td>
<td>5.961</td>
</tr>
<tr>
<td>9</td>
<td>46.3</td>
<td>9</td>
</tr>
</tbody>
</table>

Intermediate 3870
Fall 2013
### Chauvenet’s Criterion—Example 2

#### Table: Fraction of Deviation

<table>
<thead>
<tr>
<th>x</th>
<th>Frac_DEV %</th>
<th>Prob(x_i, x_mean, σ_x) = N * Prob(x_i, x_mean, σ_x)</th>
<th>Reject(x_i, x_mean, σ_x, N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>45.7</td>
<td>0.68</td>
<td>6.8</td>
</tr>
<tr>
<td>1</td>
<td>46.2</td>
<td>0.61</td>
<td>6.105</td>
</tr>
<tr>
<td>2</td>
<td>46.9</td>
<td>0.507</td>
<td>5.075</td>
</tr>
<tr>
<td>3</td>
<td>54.8</td>
<td>0.625</td>
<td>6.247</td>
</tr>
<tr>
<td>4</td>
<td>46.1</td>
<td>0.744</td>
<td>7.436</td>
</tr>
<tr>
<td>5</td>
<td>45.2</td>
<td>0.719</td>
<td>7.19</td>
</tr>
<tr>
<td>6</td>
<td>45.4</td>
<td>0.493</td>
<td>4.925</td>
</tr>
<tr>
<td>7</td>
<td>47</td>
<td>0.653</td>
<td>6.528</td>
</tr>
<tr>
<td>8</td>
<td>45.9</td>
<td>0.596</td>
<td>5.961</td>
</tr>
<tr>
<td>9</td>
<td>46.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Now recalculate the mean and standard deviation after rejecting $N_{\text{reject}}$ data points.

**Truncated data set indices and data array:**

$$j := 0 .. N - 1 - N_{\text{reject}}$$

$$X_{CH,j} := x_{\text{order}j,0}$$

**The final analysis is:**

<table>
<thead>
<tr>
<th>Including all data points</th>
<th>Excluding the rejected data point</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number data points:</td>
<td></td>
</tr>
<tr>
<td>$N = 10$</td>
<td>$N - N_{\text{reject}} = 9$</td>
</tr>
<tr>
<td>Mean:</td>
<td></td>
</tr>
<tr>
<td>$x_{\text{mean}} := \text{mean}(x) = 46.95\ m$</td>
<td>$X_{\text{mean}} := \text{mean}(X_{CH}) = 46.078$</td>
</tr>
<tr>
<td>Standard Deviation:</td>
<td></td>
</tr>
<tr>
<td>$\sigma_x := \text{stddev}(x) = 2.673\ m$</td>
<td>$\sigma_x := \text{stddev}(X_{CH}) = 0.577$</td>
</tr>
</tbody>
</table>
Intermediate Lab

PHYS 3870

Summary of Probability Theory
Probability Function (Discrete Case)

The random variable X will be called a discrete random variable if there exists a function f such that \( f(x_i) \geq 0 \) and \( \sum_i f(x_i) = 1 \) for \( i = 1, 2, 3, \ldots \) and such that for any event \( E \),

\[
P(E) = P(X \text{ is in } E) = \sum_{x \in E} f(x)
\]

where \( \sum_{x \in E} \) means sum \( f(x) \) over those values \( x_i \) that are in \( E \) and where \( f(x) = P(X = x) \).

The probability that the value of \( X \) is some real number \( x \), is given by \( f(x) = P(X = x) \), where \( f \) is called the probability function of the random variable \( X \).

Cumulative Distribution Function (Discrete Case)

The probability that the value of a random variable \( X \) is less than or equal to some real number \( x \) is defined as

\[
F(x) = P(X \leq x) = \sum f(x_i), \quad -\infty < x < \infty,
\]

where the summation extends over those values of \( i \) such that \( x_i \leq x \).

Probability Density (Continuous Case)

The random variable \( X \) will be called a continuous random variable if there exists a function \( f \) such that \( f(x) \geq 0 \) and \( \int_{-\infty}^{x} f(x) \, dx = 1 \) for all \( x \) in interval \( -\infty < x < \infty \) and such that for any event \( E \)

\[
P(E) = P(X \text{ is in } E) = \int_{x}^{E} f(x) \, dx.
\]

\( f(x) \) is called the probability density of the random variable \( X \). The probability that \( X \) assumes any given value of \( x \) is equal to zero and the probability that it assumes a value on the interval from \( a \) to \( b \), including or excluding either end point, is equal to

\[
\int_{a}^{b} f(x) \, dx.
\]
Probability Density (Continuous Case)

The random variable \( X \) will be called a continuous random variable if there exists a function \( f \) such that \( f(x) \geq 0 \) and \( \int_{-\infty}^{\infty} f(x) \, dx = 1 \) for all \( x \) in interval \( -\infty < x < \infty \) and such that for any event \( E \)

\[
P(E) = P(X \text{ is in } E) = \int_{E} f(x) \, dx.
\]

\( f(x) \) is called the probability density of the random variable \( X \). The probability that \( X \) assumes any given value of \( x \) is equal to zero and the probability that it assumes a value on the interval from \( a \) to \( b \), including or excluding either end point, is equal to

\[
\int_{a}^{b} f(x) \, dx.
\]

Cumulative Distribution Function (Continuous Case)

The probability that the value of a random variable \( X \) is less than or equal to some real number \( x \) is defined as

\[
F(x) = P(X \leq x), \quad -\infty < x < \infty
\]

\[
= \int_{-\infty}^{x} f(x) \, dx.
\]

From the cumulative distribution, the density, if it exists, can be found from

\[
f(x) = \frac{dF(x)}{dx}.
\]

From the cumulative distribution

\[
P(a \leq X \leq b) = P(X \leq b) - P(X \leq a)
\]

\[
= F(b) - F(a)
\]
Summary of Probability Theory-III

**Mathematical Expectation**

**A. Expected Value.**

Let $X$ be a random variable with density $f(x)$. Then the expected value of $X$, $E(X)$, is defined to be

$$E(X) = \sum_x x f(x)$$

if $X$ is discrete and

$$E(X) = \int_{-\infty}^{\infty} x f(x) \, dx$$

if $X$ is continuous. The expected value of a function $g$ of a random variable $X$ is defined as

$$E[g(X)] = \sum_x g(x) \cdot f(x)$$

if $X$ is discrete and

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f(x) \, dx$$

if $X$ is continuous.

**Theorems**

1. $E[aX + bY] = aE(X) + bE(Y)$
2. $E[X \cdot Y] = E(X) \cdot E(Y)$ if $X$ and $Y$ are statistically independent.
Summary of Probability Theory-IV

B. Moments

a. Moments About the Origin. The moments about the origin of a probability distribution are the expected values of the random variable which has the given distribution. The rth moment of X, usually denoted by μᵣ, is defined as

\[ μᵣ = E[Xᵣ] = \sum xᵣf(x) \]

if X is discrete and

\[ μᵣ = E[Xᵣ] = \int_{-\infty}^{\infty} xᵣf(x) \, dx \]

if X is continuous.

The first moment, μ₁, is called the mean of the random variable X and is usually denoted by μ.

b. Moments About the Mean. The rth moment about the mean, usually denoted by μᵣ, is defined as

\[ μᵣ = E[(X - μ)ᵣ] = \sum (x - μ)ᵣf(x) \]

if X is discrete and

\[ μᵣ = E[(X - μ)ᵣ] = \int_{-\infty}^{\infty} (x - μ)ᵣf(x) \, dx \]

if X is continuous.

The second moment about the mean, μ₂, is given by

\[ μ₂ = E[(X - μ)²] = μ'² - μ² \]

and is called the variance of the random variable X, and is denoted by σ². The square root of the variance, σ, is called the standard deviation.

Theorems

1. \( σ²_X = c²σ²_X \)
2. \( σ²_{X+} = σ²_X \)
3. \( σ²_{X+Y} = a²σ²_X \)