Practical Statistics

Part I - the basics

Estimators, Probability Density Functions, ChiSquare & p-value, Calibration and Simpson’s Paradox

Troels C. Petersen (Niels Bohr Institute)

“Statistics is merely a quantisation of common sense”
Practical Statistics

Part I - the basics

Estimators, Probability Density Functions, ChiSquare & p-value, Calibration and Simpson’s Paradox

“Statistics is merely a quantisation of common sense”

Troels C. Petersen (Niels Bohr Institute)
“Statistics is merely a quantisation of common sense”

Troels C. Petersen (Niels Bohr Institute)
Outline of lectures

Part I - the basics:
- Estimators
- Probability Density Functions
- ChiSquare & p-values
- Calibration
- Simpson’s Paradox

Part II - the necessities:
- Likelihood fitting
- Hypothesis testing
- Systematic uncertainties

Part III - the cool:
- Setting limits
- Look Elsewhere Effect
- The art of plotting
- The Fisher discriminant
- sPlots & sWeights
Outline of lectures

Part I - the basics:
- Estimators
- Probability Density Functions
- ChiSquare & p-values
- Calibration
- Simpson’s Paradox

Part II - the necessities:
- Likelihood fitting
- Hypothesis testing
- Systematic uncertainties

Part III - the cool:
- Setting limits
- Look Elsewhere Effect
- The art of plotting
- The Fisher discriminant
- sPlots & sWeights

Part I - the missing:
- What is probability? Axioms!
- Bayes Theorem & Jeffrey Priors
- Proof of Central Limit Theorem
- Significant digits
- Uncertainty on uncertainties

Part II - the complicated:
- Proof of Minimum Variance Bound
- Fisher Information
- Systematic uncertainty types
- Nuisance parameters

Part III - the weird:
- Details of Feldman-Cousins
- Time series
- …and surely lots more!
Why Statistics?
Why uncertainties?

In physics there are various elements of uncertainty:

- **Theory is not deterministic**
  
  Examples: Quantum effects & chaos

- **Random measurement errors**
  
  Fluctuations are present even without quantum effects!

- **Things we could know in principle but don’t…**
  
  e.g. from limitations in cost, time, etc.

We can quantify the uncertainty using **PROBABILITY**

Armed with the realisation of limitations, we can make better calculations/experiments and informed conclusions.
Example: Speed of Gravity

Imagine that you measured the speed of gravity, and got the following result:

\[ v_{\text{gravity}} = 2.89 \times 10^8 \text{ m/s} \]

That would tell you...
Example: Speed of Gravity

Imagine that you measured the speed of gravity, and got the following result:

\[ v_{\text{gravity}} = 2.89 \times 10^8 \text{ m/s} \]

That would tell you...

Nothing!!!

Because you have no idea of the uncertainty.
Example: Speed of Gravity

Imagine that you measured the speed of gravity, and got the following result:

$$v_{\text{gravity}} = 2.89 \times 10^8 \text{ m/s}$$

Depending on the uncertainty, you might foresee three very different conclusions:

$$v_{\text{gravity}} = (2.89 \pm 9.21) \times 10^8 \text{ m/s}$$  \hspace{1cm} \text{Could be anything, even negative!}

$$v_{\text{gravity}} = (2.89 \pm 0.09) \times 10^8 \text{ m/s}$$  \hspace{1cm} \text{Consistent with c, and not much else!}

$$v_{\text{gravity}} = (2.89 \pm 0.01) \times 10^8 \text{ m/s}$$  \hspace{1cm} \text{Inconsistent with c: New Discovery!!!}

(Extreme) Conclusion:

Numbers without stated uncertainties are meaningless!
Why precision?

How well do we know Newton’s Law of Gravity?

\[ F = G \frac{mM}{r^2} \]
Newton’s Law of Gravity

How well do we know Newton’s Law of Gravity? Well, reasonably well, but...

- Force central?
- Valid for all masses?
- Range of validity?
- Square Law?
- No other dependencies?
Newton’s Law of Gravity

How well do we know Newton’s Law of Gravity? Well, reasonably well, but...

- Force central?
  - Seemingly...

- Valid for all masses?
  - NO - not large ones!

- Range of validity?
  - Maybe not short ranges
  - Yes, from general relativity

- Square Law?
  - Being tested: Related to search for more dimensions

- Why is G so small?

\[ F = G \frac{mM}{r^2} \]
Why statistics in physics?

Experimental measurements are only **SAMPLES** of the reality, they can never represent the entire set of possibilities, so → they are affected by uncertainties → results can be expressed as probabilities

Theoretical calculations are mostly **APPROXIMATIONS** limited by finite resources to do the calculations or by imprecise input parameters, so → they are also affected by uncertainties → predictions can also be expressed in terms of probability

Statistics gives the understanding of **uncertainty** and **probability** in relating data and theory!!!
Why statistics in physics?

Statistics is about hypothesis testing, quantifying the answer to the question “which theory matches the data best?”

Statistics is about collecting data and logically analysing it, not being fooled by coincidences and chance observations.

Statistics is about fitting trends in data, allowing for projections and predictions.

Statistics is about understanding data, and extracting the essential information from it in the most powerful way.

Is the Higgs a spin 0 or spin 2 particle?
Biases in statistics...

When ASKING people, one may introduce (deliberate?) biases:

- **Wording 1**: Pick a color: red or blue?
- **Wording 2**: Pick a color: blue or red?

<table>
<thead>
<tr>
<th>Color Choice</th>
<th>Red</th>
<th>Blue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wording 1</td>
<td>59 %</td>
<td>41 %</td>
</tr>
<tr>
<td>Wording 2</td>
<td>45 %</td>
<td>55 %</td>
</tr>
</tbody>
</table>

One may also bias answers by giving (ir-)relevant information:

- **Wording 1**: Knowing that the population of the U.S. is 270 million, what is the population of Canada?
- **Wording 2**: Knowing that the population of Australia is 15 million, what is the population of Canada?
Biases in statistics...

When ASKING people, one may introduce (deliberate?) biases:

- **Wording 1**: Pick a color: red or blue?
- **Wording 2**: Pick a color: blue or red?

<table>
<thead>
<tr>
<th>Color Choice</th>
<th>Red</th>
<th>Blue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wording 1</td>
<td>59%</td>
<td>41%</td>
</tr>
<tr>
<td>Wording 2</td>
<td>45%</td>
<td>55%</td>
</tr>
</tbody>
</table>

One may also bias answers by giving (ir-)relevant information:

- **Wording 1**: Knowing that the population of the U.S. is 270 million, what is the population of Canada?
- **Wording 2**: Knowing that the population of Australia is 15 million, what is the population of Canada?
Mark Twain:
“There are three kinds of lies: lies, damned lies, and statistics.”

My opinion:
“The only way to convey accurate information is by statistics.”
Mark Twain:
"There are three kinds of lies: lies, damned lies, and statistics."

My opinion:
"The only way to convey accurate information is by statistics."

Hal Varian [Chief economist of Google]:
"I keep saying the sexy job in the next ten years will be statisticians."
Why statistics?
Why statistics?

Because you will need it!
(and maybe even like it)
Central Limit Theorem
Adding random numbers

If each of you chose a random number from your own favorite distribution*, and we added all these numbers, repeating this many times...

What would you expect?

* OK - to be nice to me, you agree to have similar RMSEs in these distributions!
Adding random numbers

If each of you chose a random number from your own favorite distribution*, and we added all these numbers, repeating this many times...

What would you expect?

* OK - to be nice to me, you agree to have similar RMSEs in these distributions!
Adding random numbers

If each of you chose a random number from your own favorite distribution* and we added all these numbers, repeating this many times...

What would you expect?

* OK - to be nice to me, you agree to have similar RMSEs in these distributions!

Central Limit Theorem:
The sum of N independent continuous random variables $x_i$ with means $\mu_i$ and variances $\sigma_i^2$ becomes a Gaussian random variable with mean $\mu = \Sigma_i \mu_i$ and variance $\sigma^2 = \Sigma_i \sigma_i^2$ in the limit that N approaches infinity.

OK - to be nice to me, you agree to have similar RMSEs in these distributions!
Central Limit Theorem

Central Limit Theorem:
The sum of $N$ independent continuous random variables $x_i$ with means $\mu_i$ and variances $\sigma_i^2$ becomes a Gaussian random variable with mean $\mu = \Sigma_i \mu_i$ and variance $\sigma^2 = \Sigma_i \sigma_i^2$ in the limit that $N$ approaches infinity.

The Central Limit Theorem holds under fairly general conditions, which means that the Gaussian distribution takes a central role in statistics...

The Gaussian is “the unit” of distributions!

Since measurements are often affected by many small effects, uncertainties tend to be Gaussian (until otherwise proven!).

Statistical rules often require Gaussian uncertainties, and so the central limit theorem is your new good friend.
Central Limit Theorem:

The sum of $N$ independent continuous random variables $x_i$ with means $\mu_i$ and variances $\sigma_i^2$ becomes a Gaussian random variable with mean $\mu = \Sigma_i \mu_i$ and variance $\sigma^2 = \Sigma_i \sigma_i^2$ in the limit that $N$ approaches infinity.

"The epistemological value of probability theory is based on the fact that chance phenomena, considered collectively and on a grand scale, create non-random regularity."

[Andrey Kolmogorov, Soviet mathematician, 1954]

"Nowadays, the central limit theorem is considered to be the unofficial sovereign of probability theory."

[Henk Tijms, Dutch mathematician 2004]
Example of Central Limit Theorem

Take the sum of 100 uniform numbers!
Repeat 100000 times to see what distribution the sum has…

The result is a bell shaped curve, a so-called normal or Gaussian distribution.

It turns out, that this is very general!!!
Example of Central Limit Theorem

Now take the sum of just 10 uniform numbers!
Example of Central Limit Theorem

Now take the sum of just 5 uniform numbers!
Example of Central Limit Theorem

Now take the sum of just 3 uniform numbers!

![Graph showing the distribution of the sum of 3 uniform numbers with histogram and statistical values.](image)
Example of Central Limit Theorem

This time we will try with a much more “nasty” function. Take the sum of 100 exponential numbers! Repeat 100000 times to see the sum’s distribution...

It doesn’t matter what shape the input PDF has, as long as it has finite mean and width, which all numbers from the real world has! Sum quickly becomes:

Gaussian!!!

It turns out, that this fact saves us from much trouble: Makes statistics “easy”!
Example of Central Limit Theorem

Looking at z-coordinate of tracks at vertex from proton collisions in CERNs LHC accelerator by the ATLAS detector, this is what you get:

Entries: 588400
Mean: $-4.885 \pm 0.08428$
RMS: $64.65 \pm 0.0596$
The Gaussian distribution

It is useful to know just a few of the most common Gaussian integrals:

<table>
<thead>
<tr>
<th>Range</th>
<th>Inside</th>
<th>Outside</th>
</tr>
</thead>
<tbody>
<tr>
<td>± 1σ</td>
<td>68 %</td>
<td>32 %</td>
</tr>
<tr>
<td>± 2σ</td>
<td>95 %</td>
<td>5 %</td>
</tr>
<tr>
<td>± 3σ</td>
<td>99.7 %</td>
<td>0.3 %</td>
</tr>
<tr>
<td>± 5σ</td>
<td>99.99995 %</td>
<td>0.00005 %</td>
</tr>
</tbody>
</table>
Summary

The Central Limit Theorem

...is your good friend because it...

ensures that uncertainties tend to be Gaussian

...which are the easiest to work with!
Estimators
Defining the mean

There are several ways of defining “a typical” value from a dataset:

a) Arithmetic mean  
b) Mode (most probably)  
c) Median (half below, half above)  
d) Geometric mean  
e) Harmonic mean  
f) Truncated mean (robustness)
Mean and Width

It turns out, that the best estimator for the mean is (as you all know):

$$\hat{\mu} = \frac{1}{N} \sum_i x_i = \bar{x}$$

The second (central) moment of the data is called the variance, defined as:

$$\hat{V} = \frac{1}{N} \sum_i (x_i - \mu)^2$$

Note the “hat”, which means “estimator”. It is sometimes dropped...
Mean and Width

It turns out, that the best estimator for the mean is (as you all know):

\[ \hat{\mu} = \frac{1}{N} \sum_{i} x_i = \bar{x} \]

For the standard deviation (Std), a.k.a. width or RMSE, it is:

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

Note the “hat”, which means “estimator”. It is sometimes dropped...
Mean and Width

It turns out, that the best estimator for the mean is (as you all know):

\[ \hat{\mu} = \frac{1}{N} \sum_{i} x_i = \bar{x} \]

For the standard deviation (Std), a.k.a. width or RMSE, it is:

\[ \hat{s} = \sqrt{\frac{1}{N - 1} \sum_{i} (x_i - \bar{x})^2} \]

Note the “hat”, which means “estimator”. It is sometimes dropped...
Why not “just” the naive SD?

Imagine taking 3 independent measurements, then estimating mean and SD:
Why not “just” the naive SD?

Imagine taking 3 independent measurements, then estimating mean and SD:

\[
\begin{align*}
\chi_1 & \quad \chi_2 & \quad \hat{\mu}_{\text{est.}} & \quad \mu_{\text{true}} & \quad \chi_3 \\
\hat{\sigma}_{\text{est.}} & \quad & \quad & \quad & \sigma_{\text{true}}
\end{align*}
\]

Above, all went well, because measurements were nicely distributed on both sides of the mean, and spread out according to SD.
Why not “just” the naive SD?

Imagine taking 3 independent measurements, then estimating mean and SD:

\[ X_1, X_2, X_3 \]

\[ \hat{\mu}_{\text{est.}}, \mu_{\text{true}}, \sigma_{\text{true}} \]

\[ \hat{\sigma}_{\text{est.}} \]

Above, all went well, because measurements were nicely distributed on both sides of the mean, and spread out according to SD.

\[ X_1, X_2, X_3 \]

\[ \hat{\mu}_{\text{est.}}, \mu_{\text{true}}, \sigma_{\text{true}} \]

\[ \hat{\sigma}_{\text{est.}} \]

However, now the mean is off and the Std way off (terribly so!). If we had used the true mean in the formula, it would have been less of a problem.
How incorrect is the naive SD?

Such questions can most easily be answered by a small simulation…

Produce $N=3$ numbers from a unit Gaussian, and calculate the SD estimate:

![Distribution of RMS estimates on three unit Gaussian numbers](image)

So, the “naive” SD underestimates the uncertainty significantly…
How incorrect is the naive SD?

Such questions can most easily be answered by a small simulation…
Produce N=5 numbers from a unit Gaussian, and calculate the SD estimate:

Here, the “naive” SD underestimates the uncertainty a bit…

\[ N = 5 \]
When a distribution is Gaussian, the Std. corresponds to the Gaussian width $\sigma$: 

![Gaussian distribution diagram with percentage areas labeled: 0.1% at $-3\sigma$, 2.1% at $-2\sigma$, 13.6% at $-1\sigma$, 34.1% at $-\mu$, 34.1% at $1\mu$, 13.6% at $2\sigma$, 2.1% at $3\sigma$, and 0.1% at $3\sigma$.}
Mean and Width

What is the **uncertainty on the mean**? And how quickly does it improve with more data?
Mean and Width

What is the **uncertainty on the mean**? And how quickly does it improve with more data?

\[ \hat{\sigma}_\mu = \frac{\hat{\sigma}}{\sqrt{N}} \]
Mean and Width

What is the **uncertainty on the mean**? And how quickly does it improve with more data?

$$\hat{\sigma}_\mu = \hat{\sigma} / \sqrt{N}$$

**Example:**

Cavendish Experiment
(measurement of Earth’s density)

- $N = 29$
- $\mu = 5.42$
- $\sigma = 0.333$
- $\sigma(\mu) = 0.06$

Earth density = $5.42 \pm 0.06$
Mean and Width

What is the uncertainty on the mean? And how quickly does it improve with more data?

\[ \hat{\sigma}_\mu = \frac{\hat{\sigma}}{\sqrt{N}} \]

Example:
Cavendish Experiment
(measurement of Earth’s density)

\[ N = 29 \]
\[ \mu = 5.42 \]
\[ \sigma = 0.333 \]
\[ \sigma(\mu) = 0.06 \]
Earth density = 5.42 ± 0.06
Weighted Mean

What if we are given data, which has different uncertainties? How to average these, and what is the uncertainty on the average?

\[
\hat{\mu} = \frac{\sum x_i / \sigma_i^2}{\sum 1 / \sigma_i^2}
\]

For measurements with varying uncertainty, there is no meaningful SD! The uncertainty on the mean is:

\[
\hat{\sigma}_\mu = \sqrt{\frac{1}{\sum 1 / \sigma_i^2}}
\]

Can be understood intuitively, if two persons combine 1 vs. 4 measurements
What if we are given data, which has different uncertainties?

How to average these, and what is the uncertainty on the average?

Can be understood intuitively, if two persons combine 1 vs. 4 measurements

\[
\hat{\mu} = \frac{\sum x_i}{2} \quad \hat{\mu} = \frac{1}{2} \sum \frac{1}{s_i^2}
\]

For measurements with varying uncertainty, there is no meaningful SD!

The uncertainty on the mean is:

Note that when doing a weighted mean, one should check if the measurements agree with each other!

This can be done with a ChiSquare test.

\[
\chi^2 / \text{ndf} = \frac{9.012}{8} \\
\text{Prob} = 0.3413 \\
p_0 = 3.599 \pm 0.3333
\]
Resolution using InterQuantile Range

A useful measure of resolution is the InterQuantile Range (IQR), as this is not affected by long tails.

IQR measures statistical dispersion, calculated as the difference

\[ \text{IQR} = Q_3 - Q_1 \]

The InterQuantile Efficiency (IQE) is defined as:

\[ \text{IQE} = \frac{\text{IQR}}{1.349} \]

The factor 1.349 = 2 \( \Phi^{-1}(0.75) \) ensures that IQR = 1 for a unit Gaussian.
Skewness and Kurtosis

Higher moments reveal something about a distribution's asymmetry and tails:

\[
\gamma = \frac{1}{N} \sum_{i}(x_i - \bar{x})^3 \quad \frac{1}{2} \left( \frac{1}{N} \sum_{i}(x_i - \bar{x})^2 \right)^3
\]

\[
\kappa = \frac{1}{N} \sum_{i}(x_i - \bar{x})^4 \quad \frac{1}{2} \left( \frac{1}{N} \sum_{i}(x_i - \bar{x})^2 \right)^2 - 3
\]
Correlation

Are there any correlations here?
Correlation

Are there any correlations here?

www.guessthecorrelation.com
Correlation

North Atlantic Oscillation (NAO) Effects

[Graph showing data points and correlation statistics]

www.guessthecorrelation.com
Correlation

Recall the definition of the Variance, V:

\[ V = \sigma^2 = \frac{1}{N} \sum_{i=1}^{n} (x_i - \mu)^2 = E[(x - \mu)^2] = E[x^2] - \mu^2 \]
Correlation

Recall the definition of the Variance, $V$:

$$V = \sigma^2 = \frac{1}{N} \sum_{i}^{n} (x_i - \mu)^2 = E[(x - \mu)^2] = E[x^2] - \mu^2$$

Likewise, one defines the Covariance, $V_{xy}$:

$$V_{xy} = \frac{1}{N} \sum_{i}^{n} (x_i - \mu_x)(y_i - \mu_y) = E[(x_i - \mu_x)(y_i - \mu_y)]$$
Correlation

Recall the definition of the Variance, $V$:

$$V = \sigma^2 = \frac{1}{N} \sum_{i}^{n} (x_i - \mu)^2 = E[(x - \mu)^2] = E[x^2] - \mu^2$$

Likewise, one defines the Covariance, $V_{xy}$:

$$V_{xy} = \frac{1}{N} \sum_{i}^{n} (x_i - \mu_x)(y_i - \mu_y) = E[(x_i - \mu_x)(y_i - \mu_y)]$$

“Normalising” by the widths, gives Pearson’s (linear) correlation coefficient:

$$\rho_{xy} = \frac{V_{xy}}{\sigma_x \sigma_y}$$

$$-1 < \rho_{xy} < 1$$

$$\sigma(\rho) \simeq \sqrt{\frac{1}{n} (1 - \rho^2)^2 + O(n^{-2})}$$
Correlation Matrix

The correlation matrix $V_{xy}$ explicitly looks as:

$$V_{xy} = \begin{bmatrix}
\sigma_1^2 & \sigma_{12}^2 & \cdots & \sigma_{1N}^2 \\
\sigma_{21}^2 & \sigma_{22}^2 & \cdots & \sigma_{2N}^2 \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_N^2 & \sigma_{N2}^2 & \cdots & \sigma_{NN}^2
\end{bmatrix}$$

The variance of variables can be found along the diagonal, while the (symmetric) off-diagonal terms show the co-variances.
Planck example
Correlation and Information

Correlations influence results in complex ways!

They need to be taken into account, for example in Error Propagation!

Correlations may contain a significant amount of information.

We will consider this more when we play with multivariate analysis.
Rank correlations

Sometimes, variables are perfectly correlated, just not linearly:

In this case the Pearson correlation is not the best measure.

Rank correlation compares the ranking between the two sets, and therefore gets a good measure of the correlation (see figure).

The two main cases of rank correlations are:
• Spearman’s rho
• Kendall’s tau
Rank correlations

An additional advantage is, that the rank correlation is less sensitive to outliers:

The two rank correlations are special cases of a more general rank correlation.

Typically, Spearman’s rank correlation is used.

The definition is:

\[
\rho = 1 - 6 \sum_i (r_i - s_i)^2 / (n^3 - n)
\]

where \(r_i\) and \(s_i\) is the rank of the \(i\)'th element.
Correlation

Correlations in 2D are in the Gaussian case the “degree of ovalness”!

Note how ALL of the bottom distributions have $\rho = 0$, despite obvious correlations!
Non-linear correlations (associations) are harder to measure, but possible:

- Maximal Information Coefficient (MIC), see reference and Wikipedia on MIC.
- Mutual Information (MI), linked to entropy, see Wikipedia on MI and SKLearn.
- Distance Correlation (DC) between paired vectors, see Wikipedia on DC.

Correlation Vs. Causation

“Com hoc ergo propter hoc”
(with this, therefore because of this)
Correlation Vs. Causation

“Com hoc ergo propter hoc”
(with this, therefore because of this)
Digression on correlations

Why do correlations play a fundamental role?
1. It is the fundamental relation between variables.
2. Possible independent variables give you handles (see below).
3. The degree of simplicity/linearity tells you what methods to use.
4. Correlation with variable of interest is often key.

Imagine, that you find two sets of PID variables, which are uncorrelated. In this case, you can produce two independent ways to identify signal, giving you a method for measuring performance, cross checking results, and producing enriched samples of each type. The two methods can of course be combined (with Likelihood or ML).
PDFs
Probability Density Functions
Probability Density Functions

A Probability Density Function (PDF) $f(x)$ describes the probability of an outcome $x$:

*probability to observe $x$ in the interval $[x, x+dx] = f(x) \, dx$*

PDFs are required to be normalised:

$$\int_{S} f(x) \, dx = 1$$

The expectation value (aka. mean) and the variance (i.e. standard deviation squared) are then defined as follows:

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

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 \, f(x) \, dx$$
Probability Density Functions

Example:
Consider a uniform distribution:

\[
f(x) = \begin{cases} 
1 & x \in [0, 1] \\
0 & \text{else}
\end{cases}
\]

Calculating the mean and variance:

\[
\mu = \int_{-\infty}^{\infty} x f(x) \, dx = \int_{0}^{1} x \, dx = \left[ \frac{1}{2} x^2 \right]_{0}^{1} = \frac{1}{2}
\]

\[
\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) \, dx = \int_{0}^{1} (x - \frac{1}{2})^2 \, dx = \\
\left[ \frac{1}{3} x^3 - \frac{1}{2} x^2 + \frac{1}{4} x \right]_{0}^{1} = \frac{1}{3} - \frac{1}{2} + \frac{1}{4} = \frac{1}{12}
\]
Cumulative distributions functions

Completely basic to every PDF is the cumulative distribution function, CDF, defined as:

\[ F_X(x) = \int_{-\infty}^{x} f_X(t) \, dt. \]

In words, this means that it is the probability of getting \( x \), or something below that value.

The CDF is used in many ways, and we will meet it again soon, when we discuss hypothesis testing.
Cumulative distributions functions

Completely basic to every PDF is the cumulative distribution function, CDF, defined as:

$$F_X(x) = \int_{-\infty}^{x} f_X(t) \, dt.$$ 

In words, this means that it is the probability of getting $x$, or something below that value.

The CDF is used in many ways, and we will meet it again soon, when we discuss hypothesis testing.
## Probability Density Functions

The number of PDFs is infinite, and nearly so is the list of known ones:

<table>
<thead>
<tr>
<th>Discrete distributions</th>
<th>Continuous distributions</th>
</tr>
</thead>
<tbody>
<tr>
<td>With finite support</td>
<td>Supported on a bounded interval</td>
</tr>
<tr>
<td>• Bernoulli distribution</td>
<td>• Arcsine distribution on (0,1), which is a special case of the Beta distribution on (0,1).</td>
</tr>
<tr>
<td>• Rademacher distribution</td>
<td>• Beta distribution on (0,1), of which the uniform distribution is a special case.</td>
</tr>
<tr>
<td>• Binomial distribution</td>
<td>• Lognormal distribution on (0,1).</td>
</tr>
<tr>
<td>• Beta distribution on (a,b)</td>
<td>• Dirac delta function, although not strictly a distribution, is related to it.</td>
</tr>
<tr>
<td>• Beta-binomial distribution</td>
<td>• The normal distribution is a continuous distribution.</td>
</tr>
<tr>
<td>• Degenerate distribution at ( x )</td>
<td>• Uniform distribution on (a,b) when ( a = b ).</td>
</tr>
<tr>
<td>• Binomial distribution on (a,b)</td>
<td>• Rectangular distribution is uniform on (a,b).</td>
</tr>
<tr>
<td>• Hypergeometric distribution</td>
<td>• Irwin-Hall distribution is the distribution of the sum of ( n ) independent uniform random variables.</td>
</tr>
<tr>
<td>• Poisson distribution</td>
<td>• Kert distribution on the three-dimensional sphere.</td>
</tr>
<tr>
<td>• Poisson distribution</td>
<td>• Kumaraswamy distribution is versatile.</td>
</tr>
<tr>
<td>• Wallenius’ noncentral hypergeometric distribution</td>
<td>• Logarithmic distribution (continuous).</td>
</tr>
<tr>
<td>• Bernoulli distribution</td>
<td>• PERT distribution is a special case of the beta distribution.</td>
</tr>
<tr>
<td>• Binomial distribution</td>
<td>• Raised cosine distribution on ( [\mu - \sigma, \mu + \sigma] ).</td>
</tr>
<tr>
<td>• Normal distribution</td>
<td>• Reciprocal distribution.</td>
</tr>
<tr>
<td>• Negative binomial distribution</td>
<td>• Triangular distribution on ( [a, b] ), a special case.</td>
</tr>
<tr>
<td>• Geometric distribution</td>
<td>• Truncated normal distribution on ( [a, b] ).</td>
</tr>
<tr>
<td>• Logarithmic distribution</td>
<td>• U-quadratic distribution on ( [a, b] ).</td>
</tr>
<tr>
<td>• Negative binomial distribution</td>
<td>• von Mises distribution on the circle.</td>
</tr>
<tr>
<td>• Binomial distribution</td>
<td>• von Mises-Fisher distribution on the N-dimensional sphere.</td>
</tr>
<tr>
<td>• Normal distribution</td>
<td>• Wigner semicircle distribution is important in quantum mechanics.</td>
</tr>
</tbody>
</table>

Supported on semi-infinite intervals, usually \( (0, \infty) \):

<table>
<thead>
<tr>
<th>Beta distribution</th>
<th>Beta-prime distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birnbaum-Saunders distribution</td>
<td>Chi distribution</td>
</tr>
<tr>
<td>Chi-squared distribution, also known as the noncentral chi distribution</td>
<td>Chi-squared distribution, which is the sum of ( n ) independent chi-squared random variables, which are not negative.</td>
</tr>
<tr>
<td>Exponentially modified Gaussian distribution</td>
<td>Exponentially modified Gaussian distribution</td>
</tr>
<tr>
<td>Fisher-Tippett, extreme value, or log-</td>
<td>Gumbel distribution, a special case of the Fisher-Tippett distribution.</td>
</tr>
<tr>
<td>Gumbel distribution, a special case of the Fisher-Tippett distribution.</td>
<td></td>
</tr>
</tbody>
</table>

Supported on the whole real line:

<table>
<thead>
<tr>
<th>Beta distribution</th>
<th>Beta-prime distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birnbaum-Saunders distribution</td>
<td>Chi distribution</td>
</tr>
<tr>
<td>Chi-squared distribution, also known as the noncentral chi distribution</td>
<td>Chi-squared distribution, which is the sum of ( n ) independent chi-squared random variables, which are not negative.</td>
</tr>
<tr>
<td>Exponentially modified Gaussian distribution</td>
<td>Exponentially modified Gaussian distribution</td>
</tr>
<tr>
<td>Fisher-Tippett, extreme value, or log-</td>
<td>Gumbel distribution, a special case of the Fisher-Tippett distribution.</td>
</tr>
<tr>
<td>Gumbel distribution, a special case of the Fisher-Tippett distribution.</td>
<td></td>
</tr>
</tbody>
</table>

And surely more!
Probability Density Functions

The number of PDFs is infinite, and nearly so is the list of known ones:

- The number of PDFs is infinite, and nearly so is the list of known ones:

- The Gibbs distribution
- The Maxwell–Boltzmann distribution
- The Bose distribution
- The extended negative binomial distribution
- The extended hypergeometric distribution
- The generalized log-series distribution
- The generalized normal distribution
- The geometric distribution, a discrete distribution
- The hypergeometric distribution
- The logarithmic (series) distribution
- The negative binomial distribution or Pascal distribution
- The parabolic fractal distribution
- The Poisson distribution, which describes a very l
- Poisson, the hyper-Poisson, the general Poisson t
- The Conway–Maxwell–Poisson distribution, a hi
- The Polya–Eggenberger distribution
- The Skellam distribution, the distribution of the difference
- The skew distribution
- The skew normal distribution
- The Yule–Simon distribution
- The zeta distribution has applications in statistics
- Zipf's law or the Zipf distribution. A discrete power-
- The Zipf–Mandelbrot law is a discrete power law dis

Continuous distributions

- Supported on a bounded interval
- The Arcsin distribution on [0,1], which is a spe
- The Beta distribution on (0,1), of which the uni
- The Lognormal distribution on (0,1)
- The Dirichlet function although not strictly a f
- The continuous uniform distribution on (a,b)
- The rectangular distribution is a uniform distrib
- The Inversion distribution is the distribution of t
- The Kent distribution on the three-dimensional s
- The Multivariate normal distribution is as versat
- The Multivariate Student's t distribution (Contin
- The PERT distribution is a special case of the be
- The raised cosine distribution on [μ – s, μ + s]
- The t distribution

Discrete distributions

- With finite support
- The Bernoulli distribution, which takes value 1 with
- The Rademacher distribution, which takes value 1 wi
- The binomial distribution, which describes the num
- The beta-binomial distribution, which describes the
- The degenerate distribution at x, where X is a cer
- The discrete uniform distribution, where all element
- The hypergeometric distribution, which describes t
- The Poisson binomial distribution, which describes
- Fisher's noncentral hypergeometric distribution
- Wallenius' noncentral hypergeometric distribution

And surely more!

https://docs.scipy.org/doc/scipy/reference/stats.html
Probability Density Functions

The number of PDFs is infinite, and nearly so is the list of known ones:

"Essentially, all models are wrong, but some are useful"

[George E. P. Box, British Statistician, 1919-2013]
Probability Density Functions

An almost complete list of those we will deal with in this course is:
• **Gaussian** (aka. Normal)
• **Poisson**
• **Binomial** (and also **Multinomial**)
• **Students t-distribution**
• **Uniform**
• **ChiSquare**
• **Exponential**
• **Error function** (integral of Gaussian)

You should already know most of these, and the rest will be explained.

See Barlow chap.3 and Cowan chap.2
Given \( N \) trials each with \( p \) chance of success, how many successes \( n \) should you expect in total?

This distribution is… Binomial, with

- Mean = \( Np \)
- Variance = \( Np(1-p) \)

This means, that the error on a fraction \( f = n/N \) is:

\[
\sigma(f) = \sqrt{\frac{f(1-f)}{N}}
\]
Binomial, Poisson, Gaussian

\[ f(n; N, p) = \binom{N}{n} \left(\frac{N!}{n!(N-n)!}\right) p^n (1-p)^{N-n} \]

The binomial distribution was first introduced by Jacob Bernoulli in 1713 (posthumously).

The binomial distribution basically consists of two elements: The binomial coefficient (green) and the probabilities of exactly \( n \) such events (blue).

Even though a system has many outcomes, it is typically possible to refer to either “success” of “failure”.

Assume the probability to have COVID19 is 1%. In a sample of 50 people the chance to have 1 or more infected is: \( 1-p(0) = 1 - 0.99^{50} = 0.60 \)

\[(x+y)^4 = x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + y^4.\]

\[
\begin{array}{cccccccc}
0: & 1 \\
1: & 1 & 1 \\
2: & 1 & 2 & 1 \\
3: & 1 & 3 & 3 & 1 \\
4: & 1 & 4 & 6 & 4 & 1 \\
5: & 1 & 5 & 10 & 10 & 5 & 1 \\
6: & 1 & 6 & 15 & 20 & 15 & 6 & 1 \\
7: & 1 & 7 & 21 & 35 & 35 & 21 & 7 & 1 \\
8: & 1 & 8 & 28 & 56 & 70 & 56 & 28 & 8 & 1 \\
\end{array}
\]
Binomial, Poisson, Gaussian

Requirements to be Binomial:
• Fixed number of trials, $N$
• Independent trials.
• Only two outcomes (success/failure).
• Constant probability of success/failure.

If number of possible outcomes is more than two ⇒ Multinomial distribution.

Examples of Binomial experiments:
• Tossing a coin 20 times counting number of tails.
• Asking 200 people if they watch sports on TV.
• Rolling a die to see if a 6 appear (Multinomial for all outcomes!).
• Asking 100 die-hards from Enhedslisten, if they would vote for Konservative at the next election!

Examples which aren’t Binomial experiments:
• Rolling a die until a 6 appears (not fixed number of trials).
• Drawing 5 cards for a poker hand (no replacement ⇒ not independent)
Binomial, Poisson, Gaussian

If $N \to \infty$ and $p \to 0$, but $Np \to \lambda$ then a Binomial approaches a Poisson: (see Barlow 3.3.1)

$$f(n, \lambda) = \frac{\lambda^n}{n!} e^{-\lambda}$$

In reality, the approximation is already quite good at e.g. $N=50$ and $p=0.1$.

The Poisson distribution only has one parameter, namely $\lambda$. Mean = $\lambda$ Variance = $\lambda$

So the error on a number is...

...the square root of that number!
Binomial, Poisson, Gaussian

The error on a (Poisson) number... is the square root of that number!!!
Binomial, Poisson, Gaussian

The error on a (Poisson) number... is the square root of that number!!!

A very useful case of this is the error to assign a bin in a histogram, if there is reasonable statistics \( N_i > 5-20 \) in each bin.
The error on a (Poisson) number... is the square root of that number!!!

Note: The sum of two Poissons with $\lambda_a$ and $\lambda_b$ is a new Poisson with $\lambda = \lambda_a + \lambda_b$. (See Barlow pages 33-34 for proof)
Binomial, Poisson, Gaussian

The Poisson distribution has the advantage that **neither the number of trials** $N$ 
**nor the probability of success** $p$ **has to be known** - just their product.

A typical use is when dealing with **rates** in a given interval of time, distance, 
area, volume, etc.
The Poisson distribution has the advantage that neither the number of trials $N$ nor the probability of success $p$ has to be known - just their product.

A typical use is when dealing with rates in a given interval of time, distance, area, volume, etc.

Example (real from 1898):
There were 122 deaths by horse kicks over 10 different regiments, over 20 years. What is the predicted number of deaths in a specific regiment and year?

First we estimate the mean value:

$$\mu = \frac{122}{20 \times 10} = 0.61$$

This means that the probability that 0 will die is given by:

$$P(0) = e^{0} \frac{0.61^{0}}{0!} = 0.54$$
Quick Quiz

You need to know the efficiency of your PID system for positrons.

Find 1000 data events where two electron candidates have a combined mass of 91.2 GeV (Z\(^0\)) and the negative candidate is identified as an electron (“Tag-and-probe” technique).

In 900 events the positive candidate is also identified as an electron. In 100 events it is not. Efficiency is 90%, but what about the uncertainty?

Colleague A says \(\sqrt{900} = 30\), thus 90.0 ± 3.0 %
Colleague B says \(\sqrt{100} = 10\), thus 90.0 ± 1.0 %

Which is right?
Quick Quiz

You need to know the efficiency of your PID system for positrons.

Find 1000 data events where two electron candidates have a combined mass of 91.2 GeV ($Z^0$) and the negative candidate is identified as an electron ("Tag-and-probe" technique).

In 900 events the positive candidate is also identified as an electron. In 100 events it is not. Efficiency is 90%, but what about the uncertainty?

Colleague A says $\sqrt{900} = 30$, thus $90.0 \pm 3.0\%$
Colleague B says $\sqrt{100} = 10$, thus $90.0 \pm 1.0\%$

Which is right? **Neither!**
This is not a Poisson but a Binomial ($N = 1000$ trials, $p = 0.9$ of success) Uncertainty is $\sqrt{N \cdot p \cdot (1-p)} = 9.49$, thus $90.0 \pm 0.9\%$

From previous page: $\sigma(f) = \sqrt{\frac{f(1-f)}{N}}$
Binomial, Poisson, Gaussian

If $\lambda \to \infty$, the Poisson becomes a Gaussian... ...and $\lambda > 20$ is enough!

For proof, see Barlow p.40

All fields encounter the Gaussian, and for this reason, its scale has many names!
Binomial, Poisson, Gaussian

If $\lambda \to \infty$, the Poisson becomes a Gaussian... ...and $\lambda > 20$ is enough!
Binomial, Poisson, Gaussian

If $\lambda \to \infty$, the Poisson becomes a Gaussian... ...and $\lambda > 20$ is enough!

However, note that the TAILS are not quite the same!!!
This is the very reason for the difference between Chi2 and (binned) likelihood!
“If the Greeks had known it, they would have deified it. It reigns with serenity and in complete self-effacement amid the wildest confusion. The more huge the mob and the greater the apparent anarchy, the more perfect is its sway. It is the supreme Law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshalled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to be latent all along.” [Karl Pearson]
Binomial, Poisson, Gaussian

The Gaussian defines the way we consider uncertainties.

<table>
<thead>
<tr>
<th>Range</th>
<th>Inside</th>
<th>Outside</th>
</tr>
</thead>
<tbody>
<tr>
<td>± 1σ</td>
<td>68 %</td>
<td>32 %</td>
</tr>
<tr>
<td>± 2σ</td>
<td>95 %</td>
<td>5 %</td>
</tr>
<tr>
<td>± 3σ</td>
<td>99.7 %</td>
<td>0.3 %</td>
</tr>
<tr>
<td>± 5σ</td>
<td>99.99995 %</td>
<td>0.00005 %</td>
</tr>
</tbody>
</table>
Student’s t-distribution

Given only a small (n obs.) sample (still assumed Gaussian), we don’t know the mean $\mu$ and width $\sigma$ well - we only know estimates of them! This changes the PDF to:

$$p(x | \nu, \hat{\mu}, \hat{\sigma}^2) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right) \sqrt{\pi \nu \hat{\sigma}^2}} \left(1 + \frac{1}{\nu} \left(\frac{x - \hat{\mu}}{\hat{\sigma}}\right)^2\right)^{-\frac{\nu+1}{2}}$$

$\nu = N_{\text{DoF}} = n - 1$
**Student’s t-distribution**

Given only a small (n obs.) sample (still assumed Gaussian), we don’t know the mean $\mu$ and width $\sigma$ well - we only know estimates of them! This changes the PDF to:

$$
p(x \mid \nu, \hat{\mu}, \hat{\sigma}^2) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right) \sqrt{\pi \nu \hat{\sigma}^2}} \left(1 + \frac{1}{\nu} \left(\frac{x - \hat{\mu}}{\hat{\sigma}}\right)^2\right)^{-\frac{\nu+1}{2}} \quad \nu = N_{\text{DoF}} = n - 1
$$

“Discovered” by William Gosset, student’s t-distribution takes into account the lacking knowledge of the mean and variance (as is the case for small samples).
**Student’s t-distribution**

Given only a small (n obs.) sample (still assumed Gaussian), we don’t know the mean $\mu$ and width $\sigma$ well - we only know estimates of them! This changes the PDF to:

$$p(x \mid \nu, \hat{\mu}, \hat{\sigma}^2) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2}) \sqrt{\pi \nu \hat{\sigma}^2}} \left(1 + \frac{1}{\nu} \left(\frac{x - \hat{\mu}}{\hat{\sigma}}\right)^2\right)^{-\frac{\nu+1}{2}} \nu = N_{\text{DoF}} = n - 1$$

“Discovered” by William Gosset, student’s t-distribution takes into account the lacking knowledge of the mean and variance (as is the case for small samples).

![Graphs showing Gaussian and Student's t-distributions with different degrees of freedom](image)

When mean and width are poorly known, estimating it from sample gives:

**Gaussian:** $z = \frac{x - \mu}{\sigma}$

**Student’s:** $t = \frac{x - \hat{\mu}}{\hat{\sigma}}$
I like the following overview of the most common PDFs, though it is far from perfect. However, it shows what makes the essential differences between PDFs.
Distribution Relationship

The different PDFs are related.

As can be seen, essentially all PDFs “converges” towards the Gaussian (normal) distribution.
Distribution Relationship

The different PDFs are related.

As can be seen, essentially all PDFs "converges" towards the Gaussian (normal) distribution.

Don’t worry about knowing them all…. Through a long life in statistics, I have still yet to encounter all of these in use!

Relationships among common distributions. Solid lines represent transformations and special cases, dashed lines represent limits. Adapted from Leemis (1986).
Distribution Relationship

The different PDFs are related.

As can be seen, essentially all PDFs “converges” towards the Gaussian (normal) distribution.

Don’t worry about knowing them all…. Through a long life in statistics, I have still yet to encounter all of these in use!

Relationships among common distributions. Solid lines represent transformations and special cases, dashed lines represent limits. Adapted from Leemis (1986).
A perhaps simpler overview.
The ChiSquare
The discovery of Ceres

Dwarf planet and the largest asteroid (r=487km)
On the 1st of January 1801 Giuseppe Piazzi discovered “new light” and could follow this comet/planet until 11th of February. He published the positions, but due to Ceres being behind the sun, it would be out of sight until the following winter. Following the calculations of a 24 year old mathematician/physicist, it was recovered on the 31st of December 1801 by von Zach and H. Olbers. The young man’s name was Carl Friedrich Gauss, and the method he used/invented for this was…
The discovery of Ceres

Dwarf planet and the largest asteroid (r=487km)

On the 1st of January 1801 Giuseppe Piazzi discovered “new light” and could follow this comet/planet until 11th of February. He published the positions, but due to Ceres being behind the sun, it would be out of sight until the following winter. Following the calculations of a 24 year old mathematician/physicist, it was recovered on the 31st of December 1801 by von Zach and H. Olbers. The young man’s name was Carl Friedrich Gauss, and the method he used/invented for this was...

...method of least squares!
Method of Least Squares

The problem at hand is determining the curve that best fitted data:

The “best fit” is found by minimising the sum of the squares...

Originally, uncertainties were not included (not “invented” yet!)
Method of Least Squares

The method of least squares is a standard approach to the approximate solution of **overdetermined systems**, i.e. sets of equations in which there are **more equations than unknowns**.

“Least squares” means that the overall solution minimises the **sum of the squares** of the errors made in solving every single equation.

The most important application is in **data fitting**. The best fit in the least-squares sense minimises the **sum of squared residuals**, a residual being the difference between an observed value and the fitted value provided by a model.
Method of Least Squares

The problem at hand is determining the curve that best fitted data:

Originally, uncertainties were not included (not “invented” yet!)
Method of Least Squares

Look at the figure below, and determine which curve fits best...

Illustration of Least Squares' Method

Well, what do you define as “best”? 
Method of Least Squares

Look at the figure below, and determine which curve fits best...

Illustration of Least Squares' Method

Well, what do you define as “best”? And how good is it?!?
Method of Least Squares

Look at the figure below, and determine which curve fits best...

Illustration of Least Squares' Method

Well, what do you define as “best”? And how good is it?!?
Chi-Square method

Look at the figure below, and determine which curve fits best...

Illustration of ChiSquare Method

Well, what do you define as “best”?
Defining the Chi-Square

Problem Statement: Given $N$ data points $(x, y, \sigma_y)$, adjust the parameter(s) $\theta$ of a model, such that it fits data best.

The best way to do this, given uncertainties $\sigma_i$ on $y_i$ is by minimising:

$$
\chi^2(\theta) = \sum_{i}^{N} \left( \frac{y_i - f(x_i, \theta)}{\sigma_i} \right)^2
$$

The power of this method is hard to overstate!
Not only does it provide a simple, elegant and unique way of fitting data, but more importantly it provides a goodness-of-fit measure.

This is the Chi-Square test!
Chi-Square method

Look at the figure below, and determine which curve fits best...

Illustration of ChiSquare Method

Well, what do you define as “best”? 
Chi-Square method

Look at the figure below, and determine which curve fits best...

Illustration of ChiSquare Method

Well, what do you define as “best”? The Chi2 quantifies this!
Chi-Square method

Look at the figure below, and determine which curve fits best...

Illustration of ChiSquare Method

Well, what do you define as “best”? The Chi2 quantifies this!
Chi-Square method

Look at the figure below, and determine which curve fits best...

Illustration of ChiSquare Method

Well, what do you define as “best”? The Chi2 quantifies this!
Chi-Square method

Look at the figure below, and determine which curve fits best...

Illustration of ChiSquare Method

What about now with larger errors?
Chi-Square method

Look at the figure below, and determine which curve fits best...

Illustration of ChiSquare Method

What about now with larger errors?
Chi-Square method

Look at the figure below, and determine which curve fits best...

Illustration of ChiSquare Method

With larger errors all models fit the data well.

What about now with larger errors?
Chi-Square method

Look at the figure below, and determine which curve fits best...

Illustration of ChiSquare Method

What does smaller errors do?
Chi-Square method

Look at the figure below, and determine which curve fits best...

Illustration of ChiSquare Method

What does \textbf{smaller} errors do?
**Chi-Square method**

Look at the figure below, and determine which curve fits best...

Illustration of ChiSquare Method

With smaller errors there is only ONE model that fits the data well.

What does **smaller** errors do?
Defining the Chi-Square

Problem Statement: Given N data points \((x, y, \sigma_y)\), adjust the parameter(s) \(\theta\) of a model, such that it fits data best.

The best way to do this, given uncertainties \(\sigma_i\) on \(y_i\) is by minimising:

\[
\chi^2(\theta) = \sum_{i}^{N} \frac{(y_i - f(x_i, \theta))^2}{\sigma_i^2}
\]

The power of this method is hard to overstate! Not only does it provide a simple, elegant and unique way of fitting data, but more importantly it provides a goodness-of-fit measure.

This is the Chi-Square test!
Defining the Chi-Square

Problem Statement:
Given \( N \) data points \((x, y, \sigma_y)\), adjust the parameter(s) \( \theta \) of a model, such that it fits data best.

The best way to do this, given uncertainties \( \sigma_i \) on \( y_i \) is by minimising:

\[
\chi^2(\theta) = \sum_{i=1}^{N} \frac{(y_i - f(x_i, \theta))^2}{\sigma_i^2}
\]

Note that when doing a weighted mean, one should check if the measurements agree with each other!

This can be done with a ChiSquare test.

The power of this method is hard to overstate!
Not only does it provide a simple, elegant and unique way of fitting data, but more importantly it provides a goodness-of-fit measure.

This is the Chi-Square test!
Weighted mean & ChiSquare

The weighted mean is actually an **analytical ChiSquare minimisation to a constant**. The result is the same, and one can then calculate \( \text{Prob}(\chi^2, \text{N dof}) \).

**Example:**
Data (from pendulum experiment) could be four length measurement (in mm):

\[ d : [17.8 \pm 0.5, 18.1 \pm 0.3, 17.7 \pm 0.5, 17.7 \pm 0.2] \]

The output from the above data is (many digits for **checks only**):

- Mean \( = 17.8098 \text{ mm} \)
- Error on mean \( = 0.15057 \text{ mm} \)
- ChiSquare \( = 1.28574 \)
- Ndof \( = 3 \)
- Probability \( = 0.7325213 \)

**NOTE:** This seems a very nice (and precise) result, and it may very well be. BUT, it might also be, that we all four estimated it from the same photo or similarly, which could be biased by an angled view. Then we would be fooling ourselves. We will discuss such “**systematic uncertainties**” more!
Weighted mean & ChiSquare

The weighted mean is actually an analytical ChiSquare minimisation to a constant. The result is the same, and one can then calculate Prob($\chi^2$, Ndof).

Example:
Data (from pendulum experiment) could be four length measurement (in mm):
\[ d : [17.8 \pm 0.5, 18.1 \pm 0.3, 17.7 \pm 0.5, 17.7 \pm 0.2] \]

The output from the above data is (many digits for checks only):
- Mean = 17.8098 mm
- Error on mean = 0.15057 mm
- ChiSquare = 1.28574
- Ndof = 3
- Probability = 0.7325213

\[ d = (17.81 \pm 0.15) \text{ mm} \]
\[ p(\chi^2=1.3, N_{dof}=3) = 0.73 \]

NOTE: This seems a very nice (and precise) result, and it may very well be. BUT, it might also be, that we all four estimated it from the same photo or similarly, which could be biased by an angled view. Then we would be fooling ourselves. We will discuss such “systematic uncertainties” more!
Why the ChiSquare is great

…but not its magic
Number of degrees-of-freedom

How to find/calculate the Number of degrees-of-freedom (Ndof) in a fit?
Number of degrees-of-freedom

How to find/calculate the Number of degrees-of-freedom (Ndof) in a fit?

Illustration of Number of Degrees of Freedom

\[ \chi^2 = 0.0 \]

Linear: p1
Number of degrees-of-freedom

How to find/calculate the Number of degrees-of-freedom (Ndof) in a fit?

Illustration of Number of Degrees of Freedom

This can only be done in one (unique) way:

\[ \chi^2 = 0.0 \]

Linear: p1

Ndof = 0!
Number of degrees-of-freedom

How to find/calculate the Number of degrees-of-freedom (Ndof) in a fit?

Illustration of Number of Degrees of Freedom

This can only be done in one (unique) way:

\[ \text{Ndof} = 0! \]
Number of degrees-of-freedom

How to find/calculate the Number of degrees-of-freedom (Ndof) in a fit?

![Illustration of Number of Degrees of Freedom]

- Linear: $\chi^2 = 14.5$
- Exponential: $\chi^2 = 20.4$

Now there is one point “too many”:

$\text{Ndof} = 1$
Number of degrees-of-freedom

How to find/calculate the Number of degrees-of-freedom (Ndof) in a fit?

Illustration of Number of Degrees of Freedom

Of course for pol2 the solution is still unique:

\[ \text{Ndof} = 0 \]
Number of degrees-of-freedom

The number of degrees-of-freedom, \( N_{\text{dof}} \), can be calculated as the number of points in the fit minus the number of parameters in the fit function:

\[
N_{\text{dof}} = N_{\text{data points}} - N_{\text{fit variables}}
\]

Illustration of Number of Degrees of Freedom

With 39 points and 3 parameters:

\( N_{\text{dof}} = 36 \)

Parabolic: \( p2 \)

\( \chi^2 = 30.2 \)
The Chi-Square distribution and test

The Chi-Square distribution for $N_{\text{dof}}$ degrees of freedom is the distribution of the sum of the squares of $N_{\text{dof}}$ normally distributed random variables.

The Chi-Square test consists of comparing the Chi-Square value obtained from a fit with the PDF of expected Chi-Square values. This allows the calculation of the probability of observing something with the same Chi-Square value or higher...

Rule of thumb: Chi-Square should roughly match $N_{\text{dof}}$
Chi-Square probability calculation

Given a **Chi-square value** and a **number of degrees of freedom** (Ndof), one can obtain a “goodness-of-fit”.

It is known, what Chi-square values to expect given the Ndof. One can therefore compare to this (Chi-square) distribution, and see...

*what is the probability of getting this Chi-square value or something worse, assuming this is the correct fit function!*

Example:
A fit gave the Chi-square 7.1 with 5 dof. The chance of getting this Chi-square or worse is... (reading the pink bottom curve (Ndof = k = 5) at 7.1)...
Chi-Square probability calculation

Given a Chi-square value and a number of degrees of freedom (Ndof), one can obtain a "goodness-of-fit".

It is known, what Chi-square values to expect given the Ndof. One can therefore compare to this (Chi-square) distribution, and see...

*what is the probability of getting this Chi-square value or something worse, assuming this is the correct fit function!*

Example:
A fit gave the Chi-square 7.1 with 5 dof. The chance of getting this Chi-square or worse is... (reading the pink bottom curve (Ndof = k = 5) at 7.1)... $1 - 0.78 = 22\%$
Chi-Square probability calculation

In the table below, one can get a quick estimate for low $N_{\text{dof}}$. 

<table>
<thead>
<tr>
<th>Degrees of freedom (df)</th>
<th>$\chi^2$ value $^{[16]}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.004 0.02 0.06 0.15 0.46 1.07 1.64 2.71 3.84 6.64 10.83</td>
</tr>
<tr>
<td>2</td>
<td>0.10 0.21 0.45 0.71 1.39 2.41 3.22 4.60 5.99 9.21 13.82</td>
</tr>
<tr>
<td>3</td>
<td>0.35 0.58 1.01 1.42 2.37 3.66 4.64 6.25 7.82 11.34 16.27</td>
</tr>
<tr>
<td>4</td>
<td>0.71 1.06 1.65 2.20 3.36 4.88 5.99 7.78 9.49 13.28 18.47</td>
</tr>
<tr>
<td>5</td>
<td>1.14 1.61 2.34 3.00 4.35 6.06 7.29 9.24 11.07 15.09 20.52</td>
</tr>
<tr>
<td>6</td>
<td>1.63 2.20 3.07 3.83 5.35 7.23 8.56 10.64 12.59 16.81 22.46</td>
</tr>
<tr>
<td>7</td>
<td>2.17 2.83 3.82 4.67 6.35 8.38 9.80 12.02 14.07 18.48 24.32</td>
</tr>
<tr>
<td>8</td>
<td>2.73 3.49 4.59 5.53 7.34 9.52 11.03 13.36 15.51 20.09 26.12</td>
</tr>
<tr>
<td>10</td>
<td>3.94 4.86 6.18 7.27 9.34 11.78 13.44 15.99 18.31 23.21 29.59</td>
</tr>
<tr>
<td>P value (Probability)</td>
<td>0.95 0.90 0.80 0.70 0.50 0.30 0.20 0.10 0.05 0.01 0.001</td>
</tr>
</tbody>
</table>

The table above provides $\chi^2$ values for different degrees of freedom. The $P$ values are calculated based on these $\chi^2$ values. If the $P$ value is less than 0.05, the result is considered significant.
Chi-Square probability calculation

In the table below, one can get a quick estimate for low $N_{dof}$.

<table>
<thead>
<tr>
<th>Degrees of freedom (df)</th>
<th>$\chi^2$ value $^{[16]}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.004 0.02 0.06 0.15 0.46 1.07 1.64 2.71 3.84 6.64 10.83</td>
</tr>
<tr>
<td>2</td>
<td>0.10 0.21 0.45 0.71 1.39 2.41 3.22 4.60 5.99 9.21 13.82</td>
</tr>
<tr>
<td>3</td>
<td>0.25 0.58 1.01 1.42 2.37 2.66 4.64 6.25 7.82 11.34 16.27</td>
</tr>
<tr>
<td>4</td>
<td>0.49 1.00 1.65 2.42 3.45 3.84 5.35 6.66 8.14 11.07 15.09</td>
</tr>
<tr>
<td>5</td>
<td>0.78 1.42 2.29 3.22 4.35 4.84 6.45 7.79 9.24 12.09 16.92</td>
</tr>
<tr>
<td>6</td>
<td>1.63 2.20 3.07 3.83 5.35 7.23 8.56 10.64 12.59 16.81 22.46</td>
</tr>
<tr>
<td>7</td>
<td>2.17 2.83 3.82 4.67 6.35 8.38 9.80 12.02 14.07 18.48 24.32</td>
</tr>
<tr>
<td>8</td>
<td>2.73 3.49 4.59 5.53 7.34 9.52 11.03 13.36 15.51 20.09 26.12</td>
</tr>
<tr>
<td>10</td>
<td>3.94 4.86 6.18 7.27 9.34 11.78 13.44 15.99 18.31 23.21 29.59</td>
</tr>
</tbody>
</table>

Python:

```python
chi2_prob = stats.chi2.sf(chi2_value, N_{DOF})
```

sf (survival function) = 1 - CDF

P value (Probability)

<table>
<thead>
<tr>
<th>0.95</th>
<th>0.90</th>
<th>0.80</th>
<th>0.70</th>
<th>0.50</th>
<th>0.30</th>
<th>0.20</th>
<th>0.10</th>
<th>0.05</th>
<th>0.01</th>
<th>0.001</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>
</tbody>
</table>

Non-significant | Significant
Chi-Square probability interpretation

The Chi-Square probability can **roughly** be interpreted as follows:

- If $\chi^2 / \text{Ndof} \approx 1$ or more precisely if $0.01 < p(\chi^2, \text{Ndof}) < 0.99$, then **all is good**.
- If $\chi^2 / \text{Ndof} \gg 1$ or more precisely if $p(\chi^2, \text{Ndof}) < 0.01$, then your fit is probably bad! Four potential reasons: **Hypothesis/model wrong, data is faulty, errors too small or unlucky!**
- If $\chi^2 / \text{Ndof} \ll 1$ or more precisely if $0.99 < p(\chi^2, \text{Ndof})$, then your fit is **TOO good!** Two potential reasons: **Overestimated uncertainties or lucky!**

If the statistics behind the plot is **VERY** high (greater than $10^6$), then you might have a hard time finding a model, which truly describes all the features in the plot (as now tiny effects become visible), and one hardly ever gets a good Chi-Square probability. However, in this case, one should not worry too much, unless very high precision is wanted.
The Chi-Square probability can roughly be interpreted as follows:

• If \( \frac{\chi^2}{Ndof} \approx 1 \) or more precisely if \( 0.01 < p(\chi^2,Ndof) < 0.99 \), then all is good.

• If \( \frac{\chi^2}{Ndof} \gg 1 \) or more precisely if \( p(\chi^2,Ndof) < 0.01 \), then your fit is probably bad! Four potential reasons:
  - Hypothesis/model wrong, data is faulty, errors too small or unlucky!

• If \( \frac{\chi^2}{Ndof} \ll 1 \) or more precisely if \( 0.99 < p(\chi^2,Ndof) \), then your fit is TOO good! Two potential reasons:
  - Overestimated uncertainties or lucky!

If the statistics behind the plot is VERY high (greater than \( 10^6 \)), then you might have a hard time finding a model, which truly describes all the features in the plot (as now tiny effects become visible), and one hardly ever gets a good Chi-Square probability. However, in this case, one should not worry too much, unless very high precision is wanted.

**Note:** One should only use \( \chi^2 \sim Ndof \) as a rule-of-thumb, and be cautious anyway:

- \( \text{Prob}(\chi^2=3.0, Ndof=2) = 0.223 \)
- \( \text{Prob}(\chi^2=300.0, Ndof=200) = 0.000006 \)

Always calculate and consider the probability!
Chi-Square for binned data

If the data is binned (i.e. put into a histogram), then Pearson’s Chi-square applies:

\[ \chi^2 = \sum_{i \in \text{bin}} \frac{(O_i - E_i)^2}{E_i} \]

The formula (based on Poisson statistics) is:
While Pearson’s Chi-square test is quite useful, it has some limitations, especially when some bins have low statistics.

The expected cell count ($E_i$) should not be too low. Some require 5 or more, and others require 10 or more. A common rule is 5 or more in 80% of bins, but no cells with zero expected count. When this assumption is not met, Yates’s Correction can be applied.

One alternative is to divide by $O_i$ when $O_i$ is not 0 (ROOT/Minuit).

Another alternative is the likelihood fit, which does not suffer under low statistics.

Yet, another alternative is the G-test, which is more robust at low statistics. However, I’ve never seen it in use.
Example of Chi-Square

The fact that there are several minima makes fitting difficult/uncertain!

*Always give good starting values!!!*
Why the ChiSquare is (near) magic
The uncertainty on a parameter is found where the Chi2 has increased by 1 from the minimum.
The uncertainty on a parameter is found where the Chi2 has increased by 1 from the minimum.

Example of Chi-Square

\[ \chi^2 \text{ minimum} + 1 \]

\[ \chi^2 \text{ minimum} \]
Example of Chi-Square

Uncertainties need not always be symmetric (though that is usually better!)

Asymmetric uncertainties are tricky to deal with (see later and/or Barlow).
Fitting with multiple variables, one obtains a multi-dimensional parabola. This is summarised in:
- Central fit values, $\hat{\mu}$
- Covariance matrix, $\hat{V}$

The diagonals of $V$ are the variance ($= \sigma^2$) of the fit parameters. The off-diagonals of $V$ are the co-variances (and thus correlations) between fit parameters.

You should always look at these, as they reveal a lot about your fit (see later).
Example of Chi-Square

"Fitting the Universe"
Notes on the ChiSquare method

“It was formerly the custom, and is still so in works on the theory of observations, to derive the method of least squares from certain theoretical considerations, the assumed normality of the errors of the observations being one such. It is however, more than doubtful whether the conditions for the theoretical validity of the method are realised in statistical practice, and the student would do well to regard the method as recommended chiefly by its comparative simplicity and by the fact that it has stood the test of experience”.

[G.U. Yule and M.G. Kendall 1958]
Calibration
Calibration definition

"Operation that, under specified conditions, in a first step, establishes a relation between the quantity values with measurement uncertainties provided by measurement standards and corresponding indications with associated measurement uncertainties (of the calibrated instrument or secondary standard) and, in a second step, uses this information to establish a relation for obtaining a measurement result from an indication."

[International Bureau of Weights and Measures]
Calibration definition

"Operation that, under specified conditions, in a first step, establishes a relation between the quantity values with measurement uncertainties provided by measurement standards and corresponding indications with associated measurement uncertainties (of the calibrated instrument or secondary standard) and, in a second step, uses this information to establish a relation for obtaining a measurement result from an indication.”

[International Bureau of Weights and Measures]

Personally, I would shorten this to:

“Operation that, under specified conditions:
• Establishes a relation between the quantity of interest and associated information
• Uses this information to correct/improve the estimate of the quantity of interest.”

[Shortening of the above]

Let’s have a few examples…
Calibration is many things!

Every field of science involves calibration of some kind.

Calibration of CMS calorimeter timing

Calibration target of Mars rover “Curiosity”

Calibration of concentration to fluorescence intensity (chemistry)
Calibration is many things!

Every field of science involves calibration of some kind.

Calibration of CMS calorimeter timing

Calibration target of Mars rover "Curiosity"

Calibration in this case is both correction and improvement

Calibration of concentration to fluorescence intensity (chemistry)
Though calibration spans widely, there are a few general considerations:

★ **Using control sample/group:**
  - **Purpose:** To ensure that there is not some (inherent) bias.
  - **Aim:** A good control sample is large and looks “exactly” like signal.
  - **Example:** People without “signal” disease spanning same age/lifestyles.

★ **Considering result for already well determined quantity:**
  - **Purpose:** To ensure that there is not some (inherent) bias.
  - **Aim:** A good control measurement is “easy” and well measured.
  - **Example:** Unbiased momentum resolution using particle resonances (Z).

★ **Determining relation to well measurable quantity:**
  - **Purpose:** Infer quantity in question from other sources/measurements.
  - **Aim:** If one can’t measure directly, perhaps it can be done indirectly.
  - **Example:** Measuring flow of liquid in pipe using microphone (noise!).

Each field of science have their own “tricks of the trade”, and sometimes breakthroughs and Nobel Prizes are made through calibration (length scales in the Universe, search for the ether, accurate carbon 14 dating, etc.).
Example: Carbon 14 dating

Carbon 14 dating used (and uses) samples of known age (from historical sources) to calibrate the scale and uncertainties. Tree rings have played a central role!
GPS by itself is not accurate enough for planes, but by correcting GPS position using results at known places, required accuracy can be obtained.
Example calibration

Imagine a variable, $X$, which has a peak in its spectrum, but which depends on another variable, $Y$. Variations in $Y$ “smears out” the peak in $X$, and we would therefore like to calibrate for this.

<table>
<thead>
<tr>
<th>Entries</th>
<th>50000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>99.9</td>
</tr>
<tr>
<td>RMS</td>
<td>53.79</td>
</tr>
<tr>
<td>$\chi^2$/ndf</td>
<td>264.2 / 166</td>
</tr>
<tr>
<td>Prob</td>
<td>$1.876 \times 10^{-6}$</td>
</tr>
<tr>
<td>Norm(sig)</td>
<td>$3059 \pm 182.3$</td>
</tr>
<tr>
<td>Mean(peak)</td>
<td>$100.6 \pm 0.8$</td>
</tr>
<tr>
<td>Width(peak)</td>
<td>$14.73 \pm 0.77$</td>
</tr>
<tr>
<td>Norm(bkg)</td>
<td>$231.1 \pm 1.6$</td>
</tr>
</tbody>
</table>
Imagine a variable, $X$, which has a peak in its spectrum, but which depends on another variable, $Y$. Variations in $Y$ “smears out” the peak in $X$, and we would therefore like to calibrate for this.

We therefore plot $X$ as a function of $Y$, and notice a (in this case clear) correlation between $Y$ and $X$. From this we can deduce how much the peak is shifted as a function of $Y$, and hence correct for it.

$$X_{\text{calib}} = X_{\text{meas}} + ???$$
Example calibration

Imagine a variable, $X$, which has a peak in its spectrum, but which depends on another variable, $Y$. Variations in $Y$ “smears out” the peak in $X$, and we would therefore like to calibrate for this.

We therefore plot $X$ as a function of $Y$, and notice a (in this case clear) correlation between $Y$ and $X$. From this we can deduce how much the peak is shifted as a function of $Y$, and hence correct for it. A simple inspection yields:

$$X_{\text{calib}} = X_{\text{meas}} - 40(Y - 0.5)$$
Example calibration

Applying this yields a new and (much) improved resolution of the peak in X, as would also be expected. At the same time, we can check, that now there is no dependence of the calibrated value of X on Y.

We thus conclude, that the calibration worked, and (of course) describe our calibration in the paper we publish. Note that sometimes, one needs a "control sample" for which the correct value is known through other sources.

\[ X_{\text{calib}} = X_{\text{meas}} - 40(Y - 0.5) \]
Example calibration

Q: How can we “obtain” a line at say $X=100$ to be used for calibration?

A: This you have to think AHEAD of time, i.e. when planning the experiment. It might be as simple as sticking a radioactive source down, or shining light on the instrument, or sending particles through it, but you have to consider this. Otherwise, you might have a 1.000.000$ instrument of unknown working!

We thus conclude, that the calibration worked, and (of course) describe our calibration in the paper we publish. Note that sometimes, one needs a “control sample” for which the correct value is known through other sources.

$$X_{\text{calib}} = X_{\text{meas}} - 40(Y - 0.5)$$
Simpson’s Paradox
(Really: Simpson’s “apparent” Paradox)
(if time allows)
Case: Berkeley admission

In 1973, University of California, Berkeley, were considering which of their applicants got admitted. As can be seen below, there is seemingly a bias against women, as a smaller fraction of women are admitted. Is that really the case, or is there more to the data than first glance reveals?
Case: Berkeley admission

In 1973, University of California, Berkeley, were considering which of their applicants got admitted.
As can be seen below, there is seemingly a bias against women, as a smaller fraction of women are admitted.
Is that really the case, or is there more to the data than first glance reveals?

Sex Bias in Graduate Admissions: Data from Berkeley

Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation.

P. J. Bickel, E. A. Hammel, J. W. O'Connell

Case: Berkeley admission

In 1973, University of California, Berkeley, were considering which of their applicants got admitted.

As can be seen below, there is seemingly a **bias against women**, as a smaller fraction of women are admitted.

Is that really the case, or is there more to the data than first glance reveals?

---

**Table 1. Decisions on applications to Graduate Division for fall 1973, by sex of applicant—naive aggregation. Expected frequencies are calculated from the marginal totals of the observed frequencies under the assumptions (1 and 2) given in the text. \(N = 12,763, \chi^2 = 110.8, \text{ d.f.} = 1, P = 0 \text{ (18).}**

<table>
<thead>
<tr>
<th>Applicants</th>
<th>Outcome</th>
<th>Expected</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observed</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Admit</td>
<td>Deny</td>
<td>Admit</td>
</tr>
<tr>
<td>Men</td>
<td>3738</td>
<td>4704</td>
<td>3460.7</td>
</tr>
<tr>
<td>Women</td>
<td>1494</td>
<td>2827</td>
<td>1771.3</td>
</tr>
</tbody>
</table>

---

**Sex Bias in Graduate Admissions: Data from Berkeley**

Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation.

P. J. Bickel, E. A. Hammel, J. W. O’Connell

Case: Berkeley admission

In 1973, University of California, Berkeley, were considering which of their applicants got admitted.

As can be seen below, there is seemingly a bias against women, as a smaller fraction of women are admitted.

Is that really the case, or is there more to the data than first glance reveals?

Sex Bias in Graduate Admissions: Data from Berkeley

Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation.

P. J. Bickel, E. A. Hammel, J. W. O'Connell


Table 1. Decisions on applications to Graduate Division for fall 1973, by sex of applicant—naive aggregation. Expected frequencies are calculated from the marginal totals of the observed frequencies under the assumptions (1 and 2) given in the text. \( N = 12,763 \), \( \chi^2 = 110.8 \), d.f. = 1, \( P = 0 \) (18).

<table>
<thead>
<tr>
<th>Applicants</th>
<th>Observed</th>
<th>Expected</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Admit</td>
<td>Deny</td>
<td>Admit</td>
</tr>
<tr>
<td>Men</td>
<td>3738</td>
<td>4704</td>
<td>3737 / (3738+4704) = 44.3%</td>
</tr>
<tr>
<td>Women</td>
<td>1494</td>
<td>2827</td>
<td>1494 / (1494+2827) = 34.6%</td>
</tr>
</tbody>
</table>
Case: Berkeley admission

In 1973, University of California, Berkeley, were considering which of their applicants got admitted.

As can be seen below, there is seemingly a bias against women, as a smaller fraction of women are admitted.

Is that really the case, or is there more to the data than first glance reveals?

As already noted, we are aware of the pitfalls ahead in this naive approach, but we intend to stumble into every one of them for didactic reasons.

<table>
<thead>
<tr>
<th>Applicants</th>
<th>Observed</th>
<th>Expected</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Admit</td>
<td>Deny</td>
<td>Admit</td>
</tr>
<tr>
<td>Men</td>
<td>3738</td>
<td>4704</td>
<td>3737 / (3738 + 4704) = 44.3%</td>
</tr>
<tr>
<td>Women</td>
<td>1494</td>
<td>2827</td>
<td>1494 / (1494 + 2827) = 34.6%</td>
</tr>
</tbody>
</table>
Case: Berkeley admission

Bickel et al. goes on to analyse the data further with several interesting findings:

Out of 85 departments with relevant data, a few seem to show a bias... in both directions, and mostly against men!!! What!

This seems counter intuitive to what we found to begin with. Where did the bias of 277 women less than expected go?

sex. Our computations, therefore, except where otherwise noted, will be based on the remaining 85. For a start let us identify those of the 85 with bias sufficiently large to occur by chance less than five times in a hundred. There prove to be four such departments. The deficit in the number of women admitted to these four (under the assumptions for calculating expected frequencies as given above) is 26. Looking further, we find six departments biased in the opposite direction, at the same probability levels; these account for a deficit of 64 men.
Case: Berkeley admission

Bickel et al. goes on to analyse the data further with several interesting findings:

Out of 85 departments with relevant data*, a few seem to show a bias... in both directions, and mostly against men!!! What!

This seems counter intuitive to what we found to begin with. Where did the bias of 277 women less than expected go?

*Here you should ALWAYS ask, what this involves!

In this case, 16 departments either had no women applying, or did not deny any students admission.

sex. Our computations, therefore, except where otherwise noted, will be based on the remaining 85. For a start let us identify those of the 85 with bias sufficiently large to occur by chance less than five times in a hundred. There prove to be four such departments. The deficit in the number of women admitted to these four (under the assumptions for calculating expected frequencies as given above) is 26. Looking further, we find six departments biased in the opposite direction, at the same probability levels; these account for a deficit of 64 men.
Case: Berkeley admission

In order to illustrate the point, Bickel et al. gives a hypothetical (and fun!) case:

Table 2. Admissions data by sex of applicant for two hypothetical departments. For total, $\chi^2 = 5.71$, d.f. = 1, $P = 0.19$ (one-tailed).

<table>
<thead>
<tr>
<th>Applicants</th>
<th>Observed</th>
<th>Expected</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Admit</td>
<td>Deny</td>
<td>Admit</td>
</tr>
<tr>
<td>Men</td>
<td>200</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>Women</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Department of machismatics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Men</td>
<td>50</td>
<td>100</td>
<td>50</td>
</tr>
<tr>
<td>Women</td>
<td>150</td>
<td>300</td>
<td>150</td>
</tr>
<tr>
<td>Department of social warfare</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Totals</td>
<td>250</td>
<td>300</td>
<td>229.2</td>
</tr>
<tr>
<td></td>
<td>250</td>
<td>400</td>
<td>270.8</td>
</tr>
</tbody>
</table>

The two (very hypothetical) departments are clearly very fair regarding gender, but still a difference appears between the overall resulting observation and expectation.
Case: Berkeley admission

The “apparent conclusion” (Berkeley discriminates against applications from women) is a result of Simpson’s Paradox (my text):

“The effect for group, which disappears or reverses, when considering subgroups.”

It is effects such as this, which makes statistics difficult, yet at the same time very important.
Case: Berkeley admission

The “apparent conclusion” (Berkeley discriminates against applications from women) is a result of Simpson’s Paradox (my text):

“Effect for group, which disappears or reverses, when considering subgroups”.

It is effects such as this, which makes statistics difficult, yet at the same time very important.

different degree. The proportion of women applicants tends to be high in departments that are hard to get into and low in those that are easy to get into. Moreover this phenomenon is more pronounced in departments with large numbers of applicants. Figure 1

Fig. 1. Proportion of applicants that are women plotted against proportion of applicants admitted, in 85 departments. Size of box indicates relative number of applicants to the department.
Simpson’s Paradox explained

The reason for the apparent paradox arise when frequency data is unduly given causal interpretations.

The figure on the right illustrates the “paradox” nicely.

The situation can be illustrated with 2D vectors, as shown below.

A success rate \( p/q \) (successes / attempts) can be represented by vectors with a slope. Higher slope = higher success rate.

But though B1 is steeper than L1, and B2 is steeper than L2, then B1+B1 is not as steep as L1+L2.
Summary
Summary

1. The Central Limit Theorem is your (new?) friend, as it explains why you should expect Gaussian uncertainties.

2. Estimators are given formulae that you should know in order to obtain (unbiased and efficient) estimates from data.

3. PDFs are in some sense our “model building blocks”. Most originate from given processes (that you should know), and should be used accordingly.

4. The ChiSquare is THE way to perform fits, if uncertainties are Gaussian, as it provides a crucial goodness-of-fit measure.

5. Calibration is central part of experimental physics, and requires foresight, insight, and experimental planning.

6. Always consider different types/classes separately, as this augments efficiency, and saves you from Simpson’s (apparent) paradox.