Data Analysis and Statistics



Prof Adrian Oeftiger

(adrian.oeftiger@physics.ox.ac.uk)

Based on lectures given by Dr Jenny Barnes

Lecture 2

- More about types of standard deviations
- What does $\bar{x} \pm \alpha$ mean?
- Normal distribution
- Poisson distribution
- Central Limit Theorem
- Error propagation

Standard deviation: sample and population

Often just called "Standard Deviation", you might see two versions:

***Population* standard deviation:** $\sigma_N = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$, where μ is the population mean

***Sample* standard deviation:** $\sigma_{N-1} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N - 1}}$, where $\bar{x}$ is the sample mean

Which standard deviation to use in measurements

$\Longrightarrow$ If you calculated the mean $\bar{x}$ with the same data used to calculate the standard deviation, then use the sample standard deviation σ_{N-1}!

Universal definition of ***standard error***: $\alpha = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N(N-1)}} = \frac{\sigma_{N-1}}{\sqrt{N}}$

Care needed if using built-in functions during data analysis

Different computer programmes/packages can evaluate "standard deviation" in different ways -- so do check!

- Excel has two standard deviations: `STDEV.P` and `STDEV.S` which use the definitions in this course (`.P` = population, `.S` = sample)
- For example, set of data $\{1, 2, 3, 4, 5\}$: in Python, the packages `numpy` and `statistics` can give different results for "standard deviation"

In [1]:

```python
# Programme showing the different functions for calculating standard deviation
import numpy as np
import statistics as stat

# put in data
sample = [1, 2, 3, 4, 5]

print ('using statistics:', stat.stdev(sample))
print ('using numpy:', np.std(sample))
print ('using numpy with ddof=1:', np.std(sample, ddof=1))
```

```
using statistics: 1.5811388300841898
using numpy: 1.4142135623730951
using numpy with ddof=1: 1.5811388300841898
```

What does *mean* ± *standard error* signify?

Ideally, we want to find out the true value of the thing we are trying to measure (population mean)

Population mean: $\mu \doteq \lim_{N \to \infty} \frac{\sum_{N}^{\infty} x_i}{N}$

We can never know this, but our best guess ("estimator") is the **sample mean**:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

Normal, or Gaussian, Probability Distribution

(Gaussian) normal distribution function:

$$G(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

where μ is the mean and σ is the standard deviation.

- Probability that the next measurement we make falls in the range $\mu \pm \sigma$ is then 0.68
- For $\pm 2\sigma$ prob $=$ 0.95, for $\pm 3\sigma$ prob = 0.997

Let's plot it!

In [2]:
```
def G(x, mu=0, sig=1):
    return np.exp(-(x - mu)**2 / (2 * sig**2)) / (sig * np.sqrt(2 * np.pi))
```

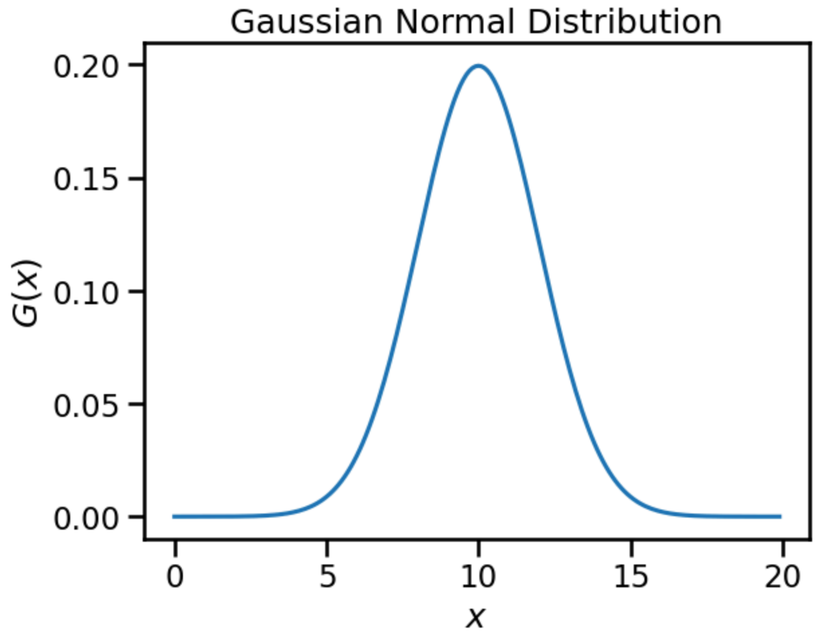
In [3]:
```
# Needed for plotting
from config import *
```

In [4]:

```
x = np.arange(0, 20, 0.1)

plt.title('Gaussian Normal Distribution')
plt.xlabel('$x$')
plt.ylabel('$G(x)$')

plt.plot(x, G(x, mu=10, sig=2));
```



Poisson distribution (1)

- Use when counting discrete events
- Conditions required:
    - counts of rare events,
    - all events are independent,
    - average rate does not change over the period of interest

Poisson distribution function for N counts:

$$P(N) = \frac{\mu^N \exp(-\mu)}{N!}$$

where μ is the mean and the standard deviation amounts to $\sqrt{\mu}$.

NB: while N is an integer, μ doesn't have to be!

Poisson distribution (2)

In [5]:
```
from scipy.special import factorial

def P(N, mu=1):
    return mu**N * np.exp(-mu) / factorial(N)
```

In [6]:
```
### Could also import Poisson distribution function from scipy library:
from scipy.stats import poisson
# P = poisson.pmf
```

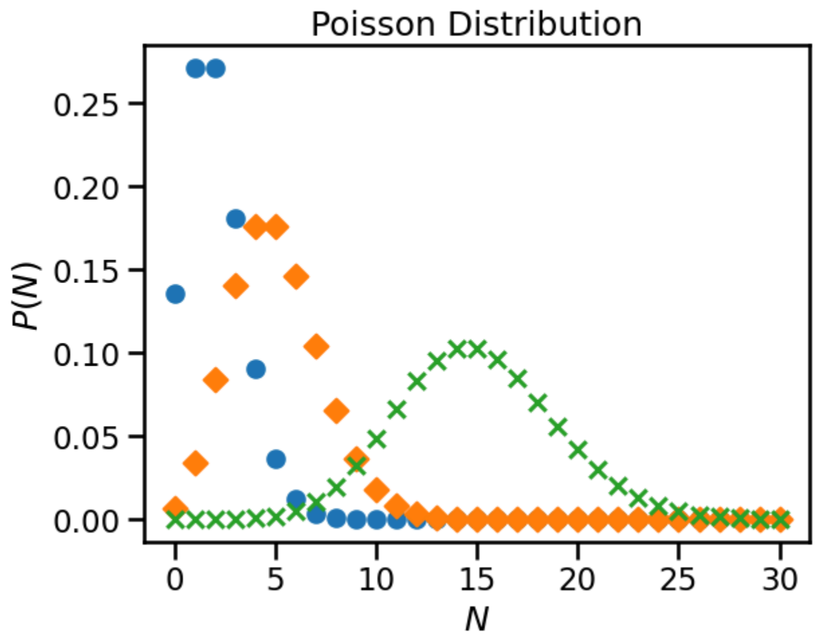
$\Longrightarrow$ try plotting for various μ:

Poisson distribution (3)

In [7]:
```
N = np.arange(0, 31, 1.)

plt.title('Poisson Distribution')
plt.xlabel('$N$')
plt.ylabel('$P(N)$')

plt.scatter(N, P(N, mu=2), marker='o');
plt.scatter(N, P(N, mu=5), marker='D');
plt.scatter(N, P(N, mu=15), marker='x');
```



Poisson distribution (3)

- Poisson distribution is *asymmetric*, especially at low values of mean
- Mean and standard deviation are related so if an experiment gives a mean count of μ, we would quote $\mu \pm \sqrt{\mu}$

At large values of mean ($\mu \gtrsim 35$), Poisson tends to Gaussian distribution (see p. 30 in Hughes and Hase):

$$P(N;\mu) \overset{\mu\to\infty}{\longrightarrow} G(N;\mu,\sigma=\sqrt{\mu})$$

Central limit theorem (1)

- Key to practical data analysis!
- The next few slides are based on Hughes and Hase, pp. 31-34,
    - please read these pages after/before the lecture

Take a sample = sequence of independent random variables x_i, drawn from **any probability distribution** which has a **finite** mean μ and standard deviation σ, then:

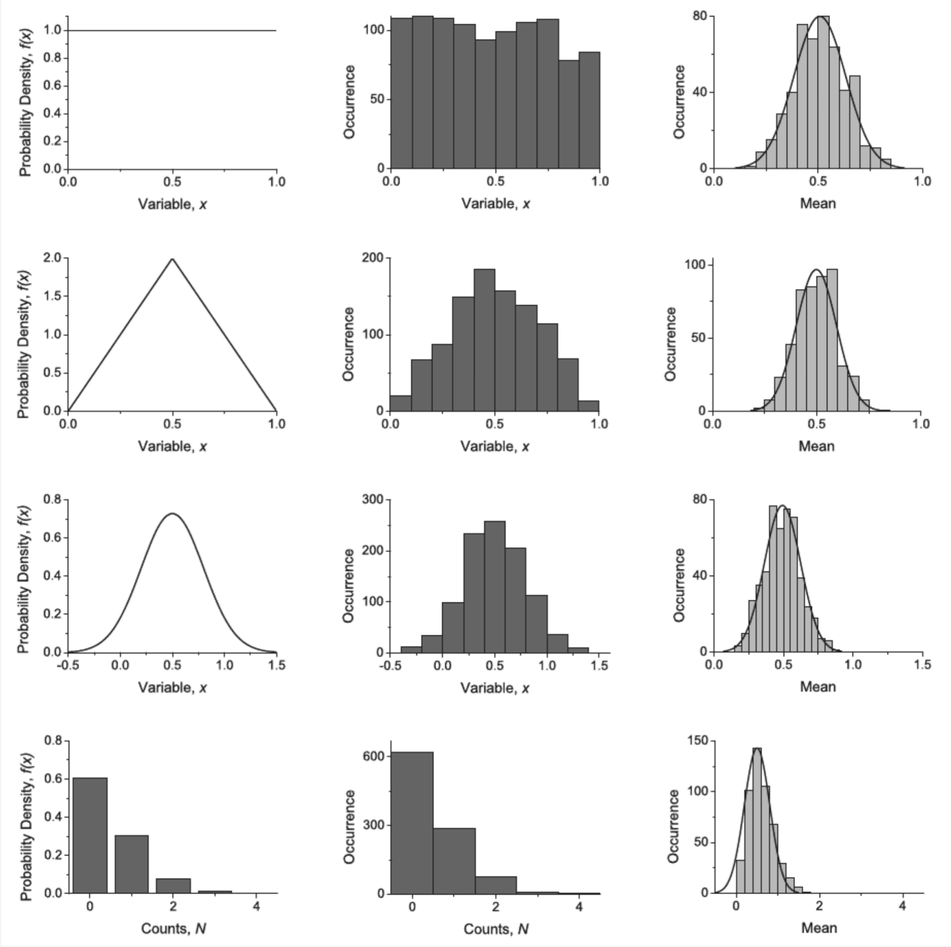
$\Longrightarrow$ The distribution of the sample mean $\bar{x} = \dfrac{x_1 + x_2 + \ldots + x_N}{N}$ is **(Gaussian) normal**, with a mean $\bar{x}$

and standard deviation $\dfrac{\sigma}{\sqrt{N}}$!

Examples central limit theorem

Taken from Hughes and Hase, p. 32:

- First column shows four different prob. distribution functions, each with mean of 0.5
    - NB: last is Poisson with a mean of the number of counts = 0.5
- Second column shows outcome of 1000 computer generated trials from the prob. distribution function (matches first column)
- Third column shows "experiments": 5 numbers chosen from prob. distribution and averaged, repeated 1000 times
    - cf. Gaussian (curve)
    - Standard deviation is smaller by $\sqrt{5}$ cf. parent standard deviation as 5 points used for each average



Central limit theorem (2)

- Reiterate: If we have a sequence of independent random variables x_i drawn from any probability distribution which has a finite mean μ and standard deviation σ, then the distribution of the sample mean $\bar{x} = \frac{x_1 + x_2 + \ldots + x_N}{N}$ is a normal probability distribution
    - with a mean $\bar{x}$ and standard deviation $\frac{\sigma}{\sqrt{N}}$
- This lets us write our value as $\bar{x} \pm \alpha$ (and why we need $\frac{1}{\sqrt{N}}$ in our standard error)
- $\bar{x} \pm \alpha$ means there is a 68% chance the next measurement x_{N+1} lies in this range

Central limit theorem (3)

Is the central limit theorem always valid?

- Central limit theorem fails if the data comes from a probability distribution which doesn't have a finite mean and/or variance, e.g. Cauchy distributed random variables or Lorentz distribution (defined by peak position and half width, doesn't integrate to 1 in general)
- Most of the time it works (most of the time you are taking measurements in the lab, it will work)

Revisiting standard error visualisation (1)

In [8]:
```
# Programme illustrating the standard error and standard deviation
import numpy as np

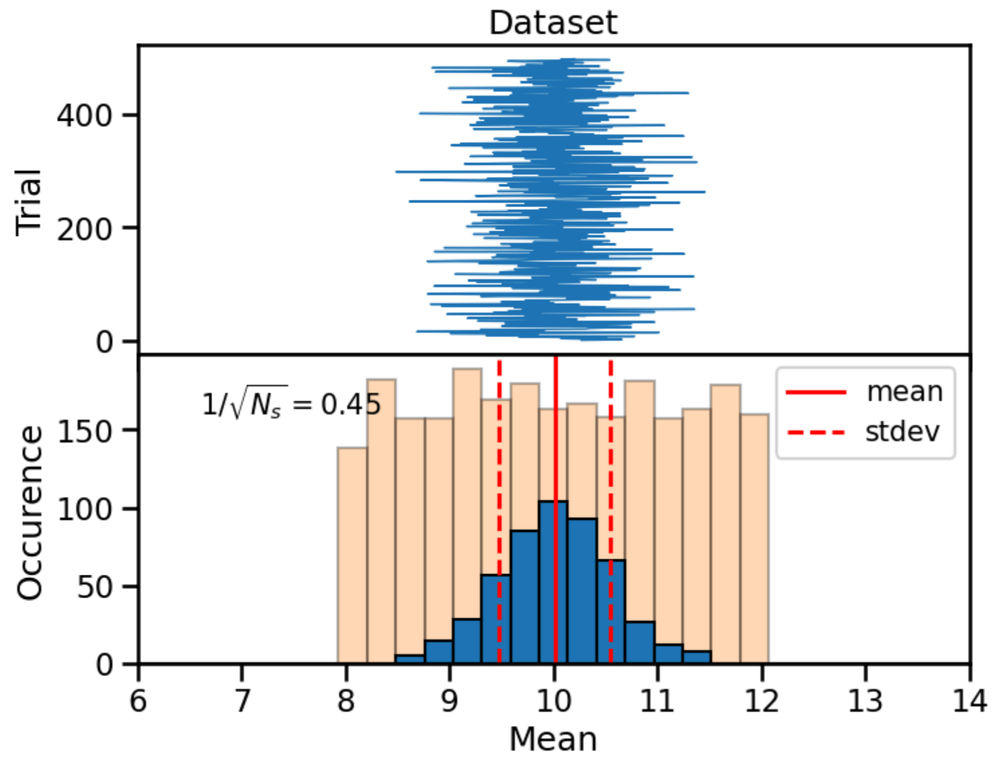
# initialise random number generator:
np.random.seed(0)

# create dataset by drawing from a random distribution:
# dataset = np.random.normal(10, 1, 2500)
# dataset = np.random.triangular(6, 10, 14, 2500)
dataset = np.random.uniform(8, 12, 2500)
# dataset = poisson.rvs(mu=10, size=2500)
```

$\Longrightarrow$ Try replacing the Gaussian (`np.random.normal`) by Poisson (`poisson.rvs`), uniform (`np.random.uniform`) or the triangular distribution (`np.random.triangular`)!

Revisiting standard error visualisation (2)

In [9]:
```
plot_dataset(dataset, N_per_sample=5);
```



Revisiting standard error visualisation (3)

We note, the distribution of means is Gaussian normal in each case, independent of the distributions behind the raw data set.

Error propagation: single variable

- We have a measured value with an error ($\bar{x} \pm \alpha$) and need to use this in a formula $f(x)$ to work out a dependent quantity
- The result is simply given by passing the mean value to the formula $f(\bar{x})$, but what about the error?

Functional approach: You explicitly calculate $f(\bar{x} - \alpha)$ and $f(\bar{x} + \alpha)$

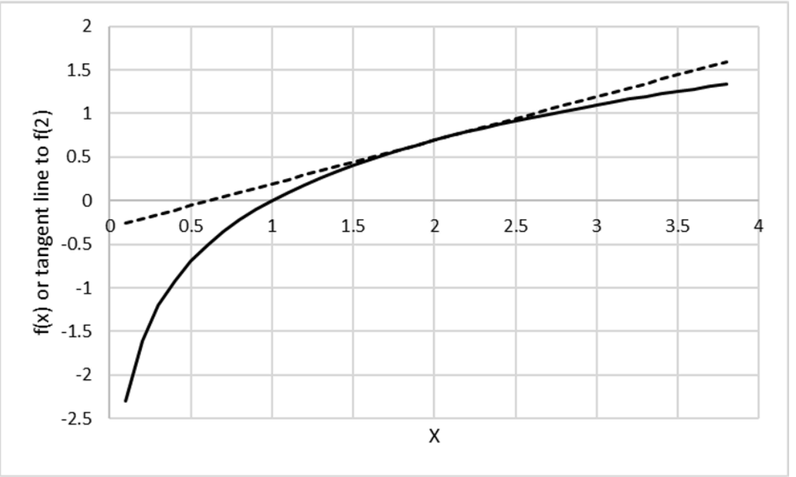
- Define $\delta f_+ \doteq f(\bar{x} + \alpha) - f(\bar{x})$ and $\delta f_- \doteq f(\bar{x}) - f(\bar{x} - \alpha)$
    - (applies for a monotonically increasing f)
- The result is then $f(\bar{x})^{\delta f_+}_{\delta f_-}$ -- e.g. $\left(2.0^{+0.5}_{-0.4}\right)$

Error propagation: calculus (1)

Alternative: **calculus method** (for small errors)

Approximate function as a straight line around $f(\bar{x})$:

$$\Longrightarrow \quad \delta f = \left| \left(\frac{df}{dx} \right)_{x=\bar{x}} \right| \cdot \alpha$$



Error propagation: calculus (2)

- The result is then $f(\bar{x}) \pm \delta f$
- Care needed: we are approximating the function as a straight line around $x = \bar{x}$ and for some functions this isn't a good idea
- The functional approach can reveal asymmetries in the error but the calculus method won’t
- Look up tables exist for functions (e.g. p. 39 in Hughes and Hase)

Error propagation: multi variables

- Example: We have a formulae for moment of inertia I of a disk: $I = \frac{1}{2}mr^2$
- We have measured the mass m and the radius r of the disk, both have a mean value and errors
- What is best estimate of I? (See this in GP08 -- rotational dynamics)
- We need some assumptions to make life easier:
    - the different variables are **independent**
    - the errors in the different variables are **uncorrelated**
- This basically means that changes in one variable don't affect the others
- If they are correlated you need to use covariance matrices (see next year!)

Take two variables...

- Suppose we had $z = f(x, y)$
- For one variable we had $\delta f = \left| \left(\frac{df}{dx} \right)_{x=\bar{x}} \right| \cdot \alpha$
- Now $\delta f_x = \left| \left(\frac{\partial f}{\partial x} \right)_{x=\bar{x}} \right| \cdot \alpha_x$ and $\delta f_y = \left| \left(\frac{\partial f}{\partial y} \right)_{y=\bar{y}} \right| \cdot \alpha_y$
- Use Pythagoras' theorem: $(\delta z)^2 = (\delta f_x)^2 + (\delta f_y)^2$
- Could extend to more variables -- known as **summing in quadrature**
- Remember that we did assume we could approximate function as a straight line and uncorrelated so *make sure these are valid*

Example: Let's do this for the moment of inertia formula

- Moment of inertia of a disk: $I = \frac{1}{2}mr^2$ with $\bar{m} \pm \alpha_m$ and $\bar{r} \pm \alpha_r$
- Value of function using mean values: $\bar{I} = \frac{1}{2}\bar{m}\bar{r}^2$
- Differentiate with respect to m or r only (keep other one constant):

$$\frac{\partial I}{\partial m} = \frac{1}{2}r^2 \qquad \text{and} \qquad \frac{\partial I}{\partial r} = mr$$

- Error in moment of inertia would be $(\delta I)^2 = \left(\frac{1}{2}\bar{r}^2\alpha_m\right)^2 + (\bar{m}\bar{r}\alpha_r)^2$
- $\Longrightarrow$ Quote value as $I \pm \delta I$
- Good to double check for your differentiating -- do you get the *correct units in each bracket*?

Multi-variables: Calculus method

- Alternatively, you can work out the fractional errors using a look up table (e.g. p. 44 in Hughes and Hase)
- In the disk case:

$$\left(\frac{\delta I}{\bar{I}}\right)^2 = \left(\frac{\delta m}{\bar{m}}\right)^2 + \left(2 \cdot \frac{\delta r}{\bar{r}}\right)^2$$

- This can also be derived from $(\delta I)^2 = \left(\frac{1}{2}\bar{r}^2 \alpha_m\right)^2 + (\bar{m}\bar{r}\alpha_r)^2$ and $\bar{I} = \frac{1}{2}\bar{m}\bar{r}^2$
- The factor of 2 appears in the last term because the original equation $I = \frac{1}{2}mr^2$ had *r* squared.

A reminder about dominant errors

- We often add errors in quadrature
- If one term is much larger than the others it will dominate
- It might be an error α is the largest or the term is raised to a power in the formula so the derivative is large, e.g.

$$(\delta I)^2 = \left(\frac{1}{2}\bar{r}^2\alpha_m\right)^2 + (\bar{m}\bar{r}\alpha_r)^2$$

$$\text{or} \qquad \left(\frac{\delta I}{\bar{I}}\right)^2 = \left(\frac{\alpha_m}{\bar{m}}\right)^2 + \left(2\cdot\frac{\alpha_r}{\bar{r}}\right)^2$$

- Here the fractional error ($\alpha_r/\bar{r}$) in the radius has more effect than the fractional error in the mass ($\alpha_m/\bar{m}$) because the formula had r^2

Summary of lecture 2

- More about types of standard deviations
- What does $\bar{x} \pm \alpha$ mean?
- Normal distribution
- Poisson distribution
- Central Limit Theorem
- Error propagation (functional vs calculus approach)

Questions?