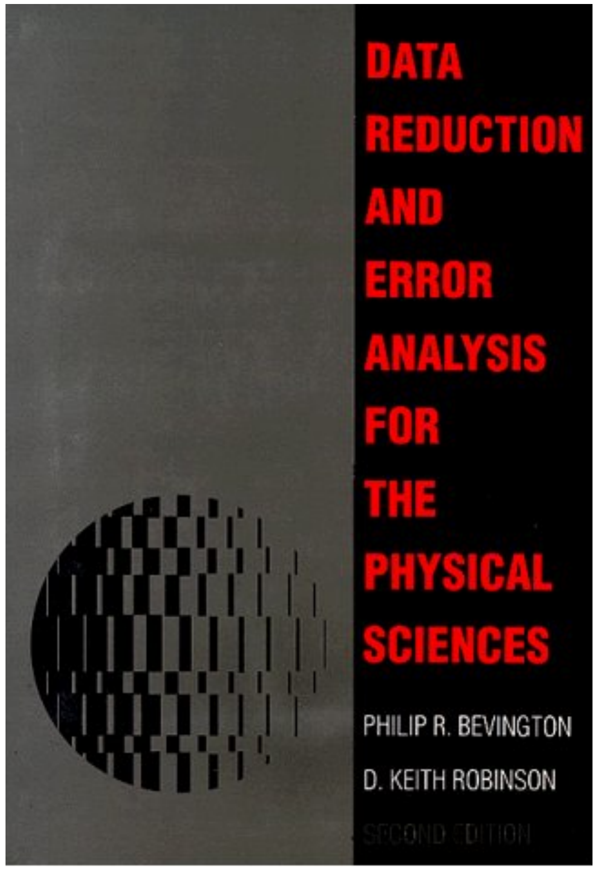
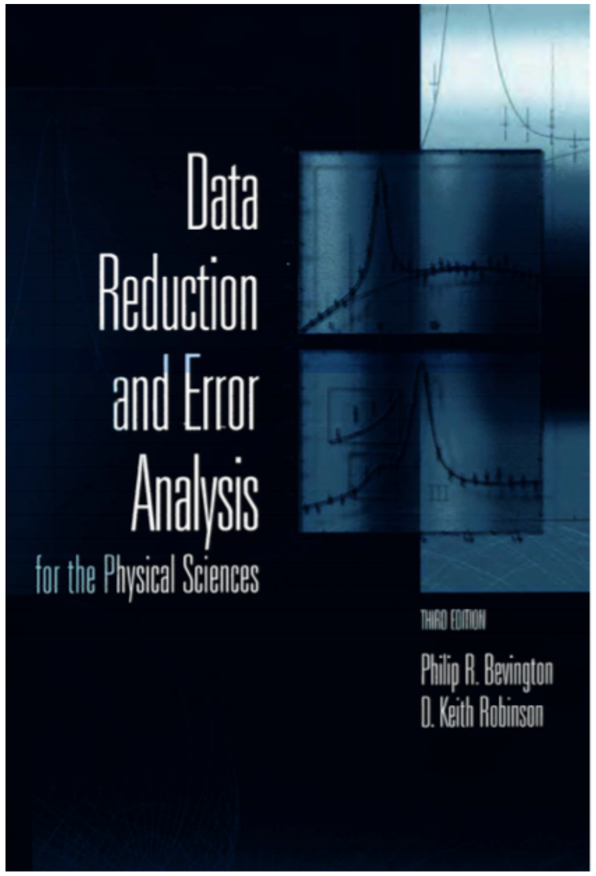
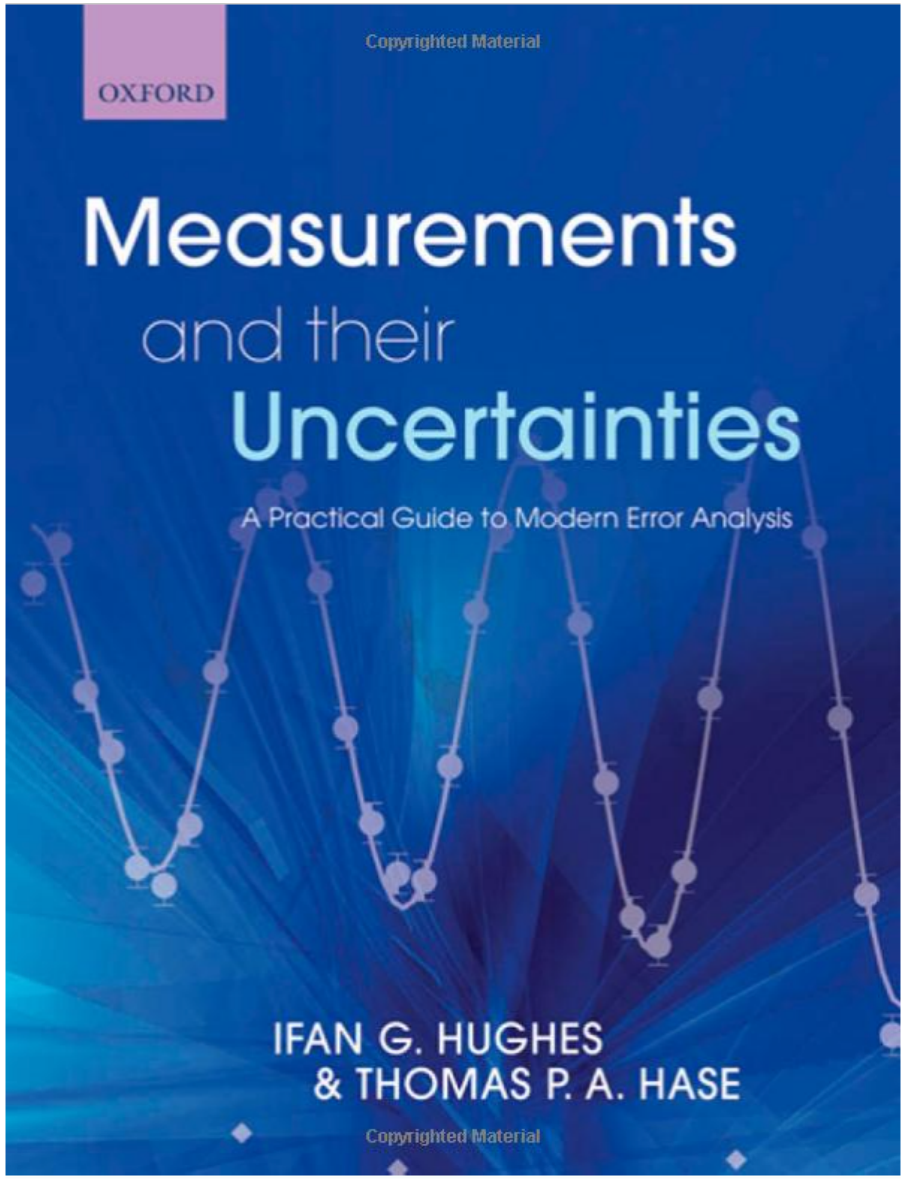
Measurement Probability Distributions

Assoc. Prof. John Quinn

Some Recommended Books







UCD Online:
http://lib.myilibrary.com/Open.aspx?id=273234#

Measurement Errors & Distributions

- Statistical (or Random) Errors:
    - we cannot measure any physical quantity with infinite precision - there is always some uncertainty on a measurement.
    - when an experiment is repeated several times we find that we do not get the exact same answer each time but that the values fluctuate about some mean.

Measurement Errors & Distributions

- Instrumentation Precision:
  - Analogue: The statistical error associated with analogue instrumentation is often due to how well one can read the scale, and it is often up to the experimenter to estimate.
  - Digital Meter: Guide: the precision of a digital meter is limited to the last digit.
    - e.g. repeated measurements of a voltage with a digital multimeter gives 8.41 V. We would thus quote the voltage as 8.41 ± 0.01 V.
  - ADC: resolution limited by number of bits, i.e. range divided by 2^N where N is the number of bits.
  - Other: instrumentation documentation
- Non-instrumental:
  - Environmental etc, beyond precision of instrument.

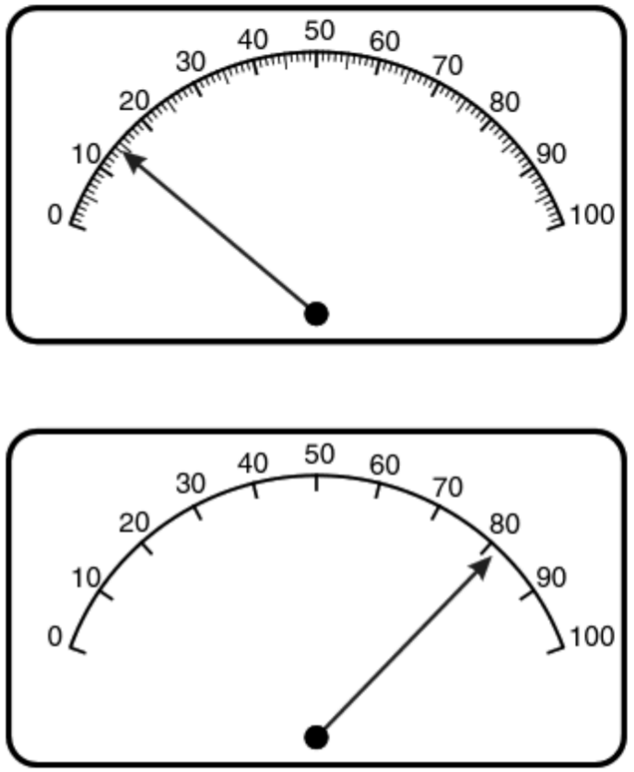


Fig. 1.4 The upper part of the figure displays a situation where estimating the uncertainty to be half a division is appropriate; in contrast in the lower part of the figure the uncertainty in the measurement is substantially smaller than half a division.

(from Hughes & Hase book)

Measurement Errors & Distributions

- Measurement Probability Distributions
  - We assume statistical errors follow a probability distribution (usually Gaussian)
  - If we know the distribution (ie. the parameters) then we can estimate the most likely value of a parameter and give a confidence interval on that estimate.
  - However, we more-than-likely do not know the true parameters which determine the distribution and these are often derived from the measurements themselves, or properties of the instrumentation.
  - The real world is complicated!
    - <u>All models are wrong, some are useful!</u>

Statistical and Systematic Errors

- Systematic and Non-Statistical Errors are not random fluctuations but additional uncertainties due to incomplete/imperfect/incorrect knowledge of experiment/calibration etc.
  - Not easy to detect and correct.
  - Generally lumped together into the term "Systematics"

- In general when we quote errors on a quantity they are the Statistical Errors.
  - Non-Statistical/Systematic Errors may be quoted in addition:
    - $x=1.0 \pm 0.1_{stat} \pm 0.2_{sys}$

- Mistakes!
  - Mistakes happen including instrument malfunction and human error.
  - Can be difficult to catch
  - Sometimes with catastrophic results.

Probability Distributions

- There are lots of probability distribution functions!
  - See, e.g.: https://docs.scipy.org/doc/scipy/reference/stats.html

- **Common probability distributions in Experimental Physics**:
  - Normal/Gaussian distribution (continuous):
    - errors generally assumed to follow
  - Student's t (continuous):
    - comparing measured mean to expected using measured standard deviation
  - Poisson (discrete):
    - number of discrete events in an interval
  - Binomial Distribution (discrete)
  - $\chi 2$ (continuous):
    - distribution of sum of normally-distributed deviates squared
    - used to test agreement between data and model or different data sets.

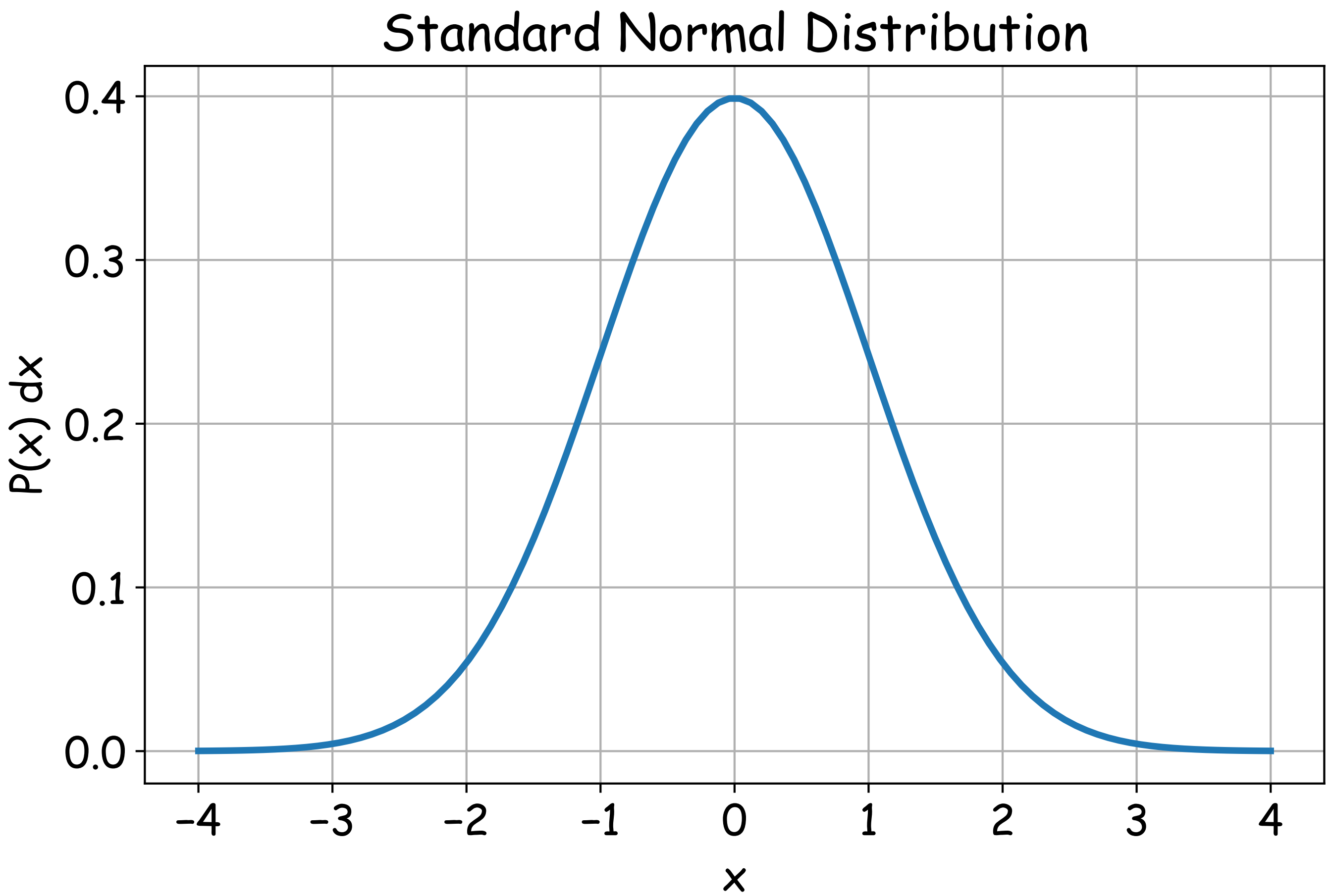
The Gaussian (Normal) Distribution

- Statistical Errors tend to follow a Gaussian Distribution

$$P_{DF}(x)\,dx = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\, dx$$

where:

- $P_{DF}(x)\,dx$ is the probability of obtaining a value between x and $x+dx$
- μ is the mean (centre) of the distribution
- σ is the standard deviation and characterises the width of the distribution
- normalised: area =1.

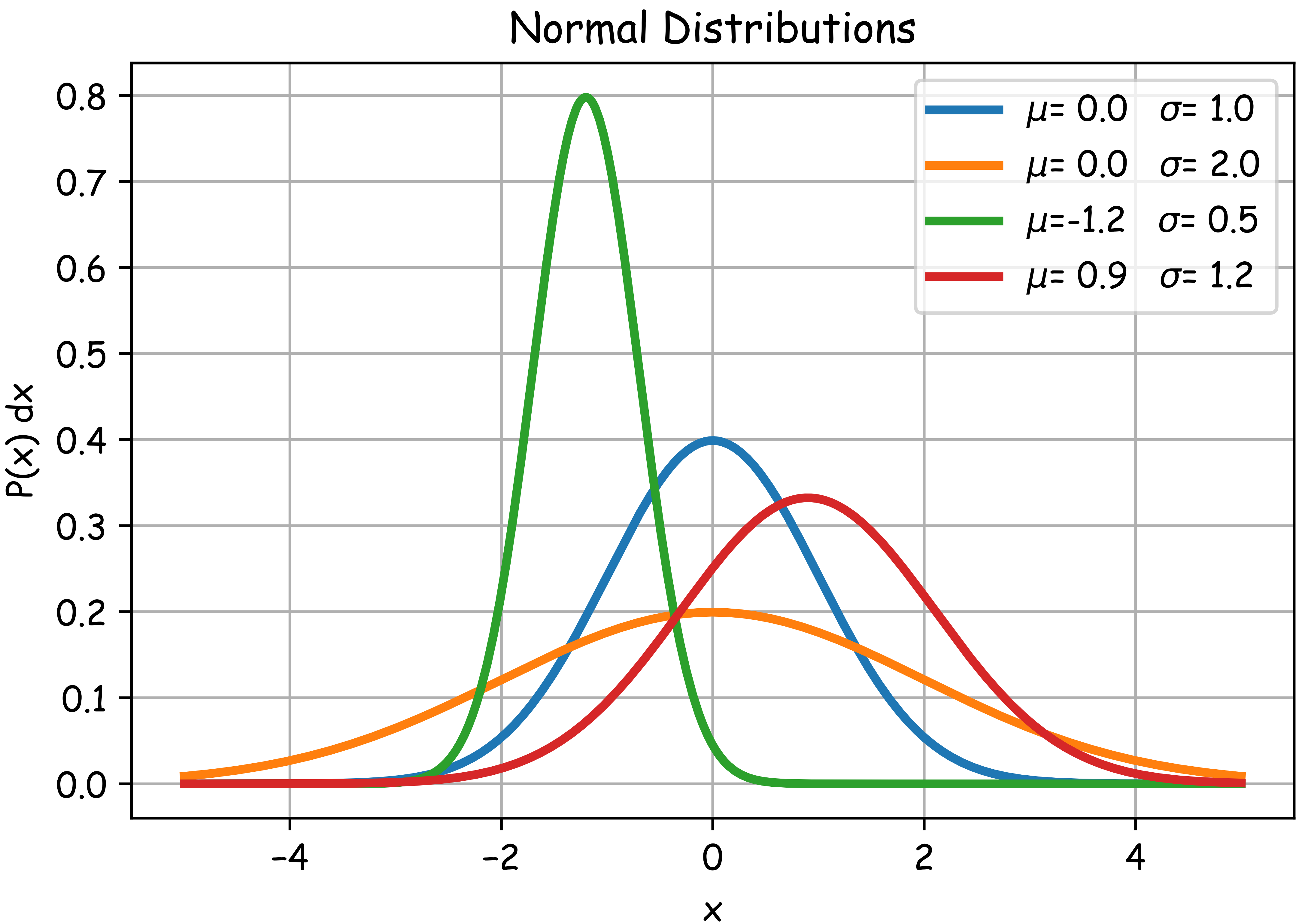


Standard Normal: $\mu=0$, $\sigma=1$

Note: for data: $\mu = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$

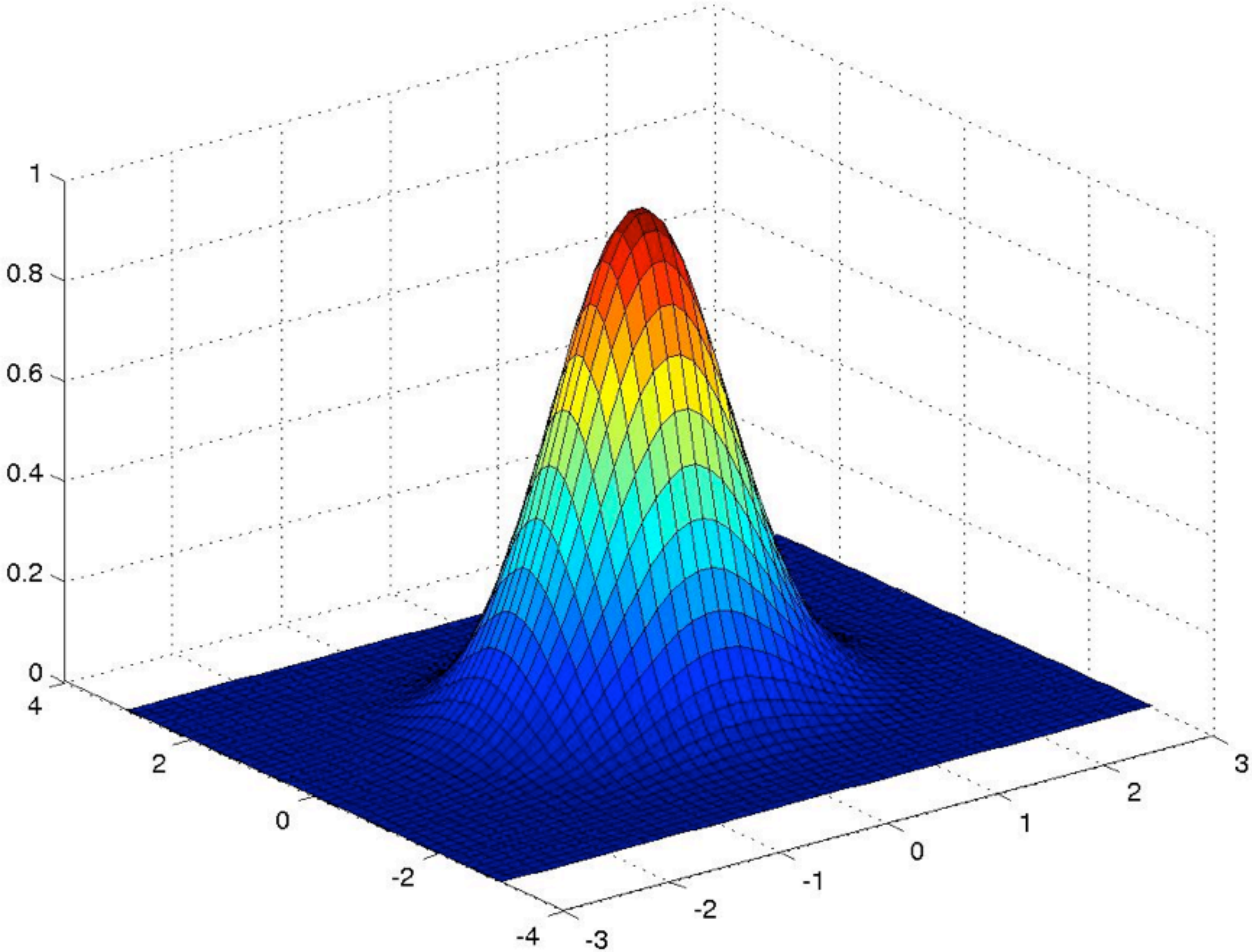
(if deriving the mean and the variance from the data)

The Gaussian/Normal Distribution



The Gaussian/Normal Distribution

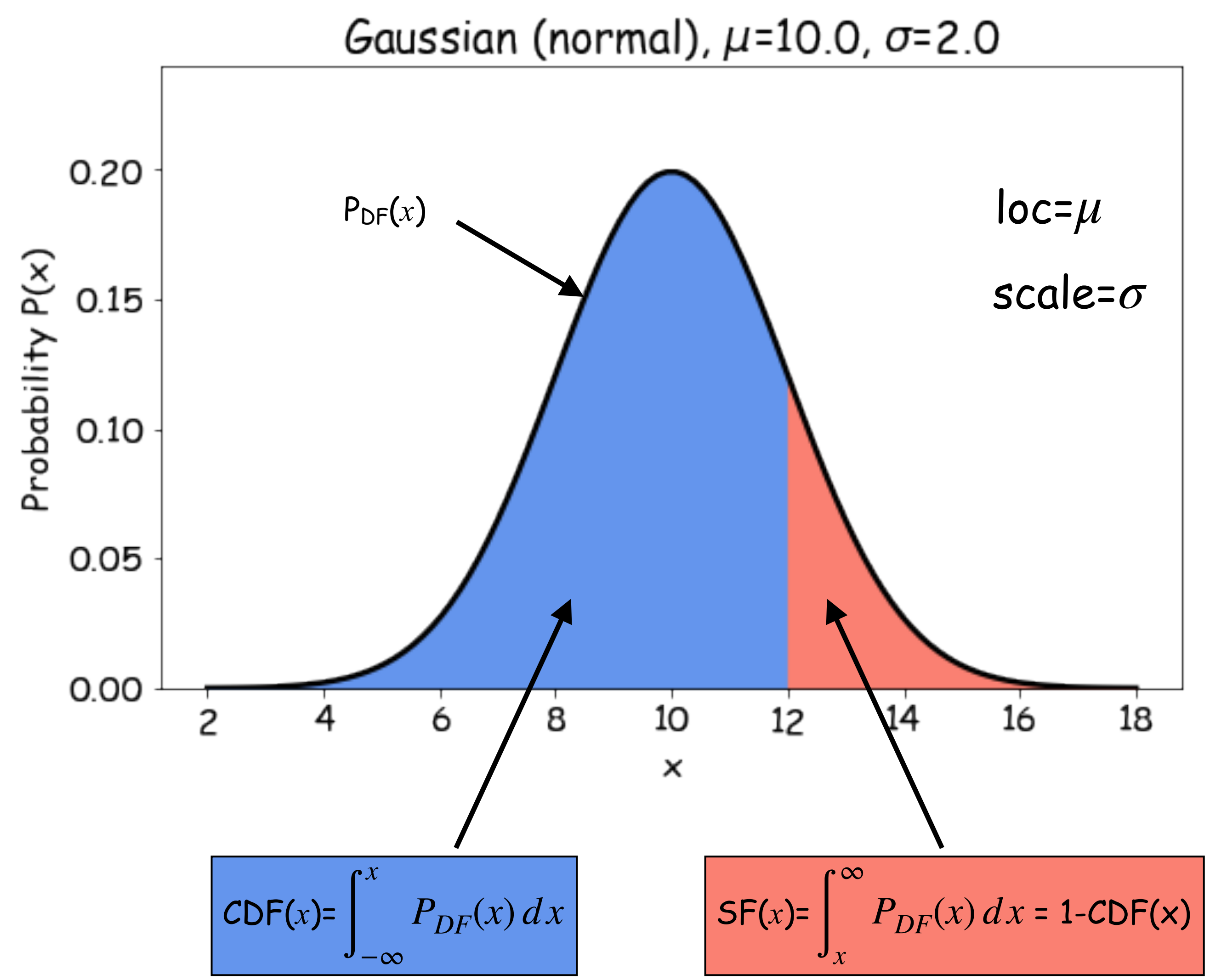
Two-dimensional Gaussian:



Scipy.stats Distributions

- Scipy.stats contains many statistical functions with a common naming structure for the member functions (inheritance).
- e.g. `scipy.stats.norm`: (some of member functions):

`rvs(loc=0, scale=1, size)`	random variates
`pdf(x, loc=0, scale=1)`	probability density function*
`cdf(x, loc=0, scale=1)`	cumulative distribution function (CDF)
`ppf(q, loc=0, scale=1)`	percent point function (inverse of CDF)
`sf(x, loc=0, scale=1)`	survival function (SF)
`isf(q, loc=0, scale=1)`	inverse survival function
`expect(fun, args, …)`	expectation value of a function
`interval(alpha, loc=0, scale=1)`	Endpoints of the range around the median that contains alpha percent of the distribution



* for discrete distributions `pdf()` is replaced with `pmf()`

Example: scipy.stats.norm()

- Note: there are two ways to use:

```
from scipy.stats import norm

mean=10
std=5

x=8
p=norm.pdf(x, loc=mean, scale=std)
cd=norm.cdf(x, loc=mean, scale=std)
sf=norm.sf(x, loc=mean, scale=std)
isf=norm.isf(sf, loc=mean, scale=std)
```

or

```
from scipy.stats import norm

mean=10
std=5

# make an instance of the class
# with specified mean and std

norm1=norm(loc=mean, scale=std)

x=8
p=norm1.pdf(x)
cd=norm1.cdf(x)
sf=norm1.sf(x)
isf=norm1.isf(sf)
```

Example 3.2.2 Hughes and Hase

- A box contains 100 Ω resistors that are known to have a standard deviation of 2 Ω.
  - (A): What is the probability of selecting a resistor of value 95 Ω or less?
  - (B): What is the probability of selecting a resistor in the range 99-101 Ω?

- (A): $P(R \leq 95\,\Omega) = \int_{-\infty}^{95} N(x; \mu = 100, \sigma = 2)\, dx$

```
norm.cdf(95, loc=100, scale=2)
0.0062
```

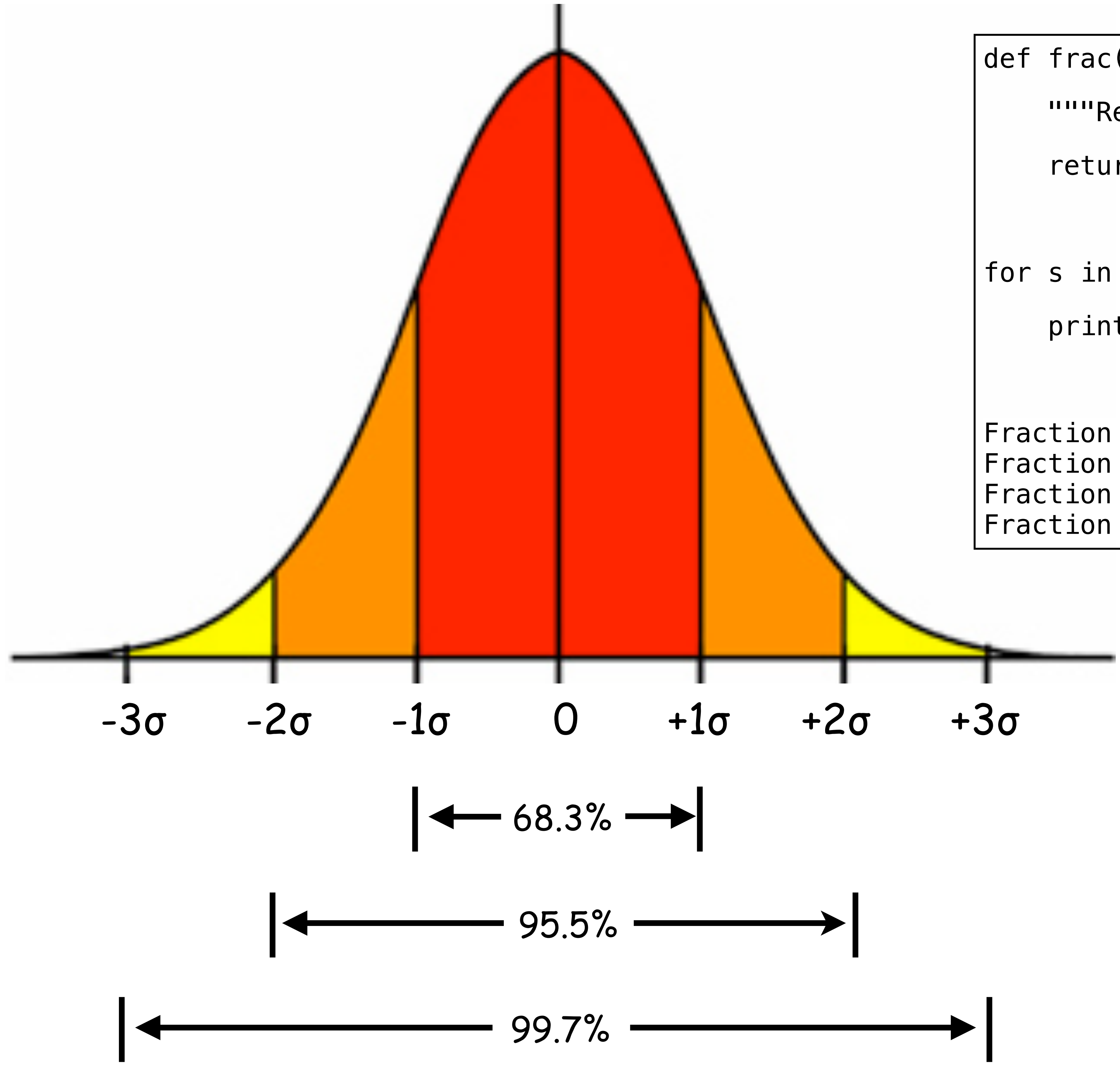
$N(x; \mu, \sigma)$ is the Normal PDF with mean μ and standard deviation σ

- (B): $P(99 \leq R \leq 101\,\Omega) = \int_{99}^{101} N(x; \mu = 100, \sigma = 2)\, dx$

```
norm1=norm(loc=100, scale=2)
norm1.cdf(101) - norm1.cdf(99)
0.38
```

The Gaussian (Normal) Distribution

- If we know the mean and standard deviation for a set of measurements (or technique), what is the probability that a measurement will fall in a given range?




```
def frac(sigma):
    """Return fraction of norm within ± sigma"""
    return norm.cdf(sigma)-norm.cdf(-sigma)

for s in range(1,5):
    print(f"Fraction within {s} sigma: {frac(s):.5f}")

Fraction within 1 sigma: 0.68269
Fraction within 2 sigma: 0.95450
Fraction within 3 sigma: 0.99730
Fraction within 4 sigma: 0.99994
```

~1 in 3 outside of range

~1 in 20 outside of range

~1 in 400 outside of range

Parent and Sample Distributions

- In general we do not know the true (parent) distribution from which our measurements are drawn and we must estimate those from the data (the sample)

$$\mu = \lim_{N\to\infty} \frac{1}{N}\sum_i x_i \qquad \sigma_{parent} = \lim_{N\to\infty} \sqrt{\frac{\sum (x_i - \mu)^2}{N_{parent}}}$$

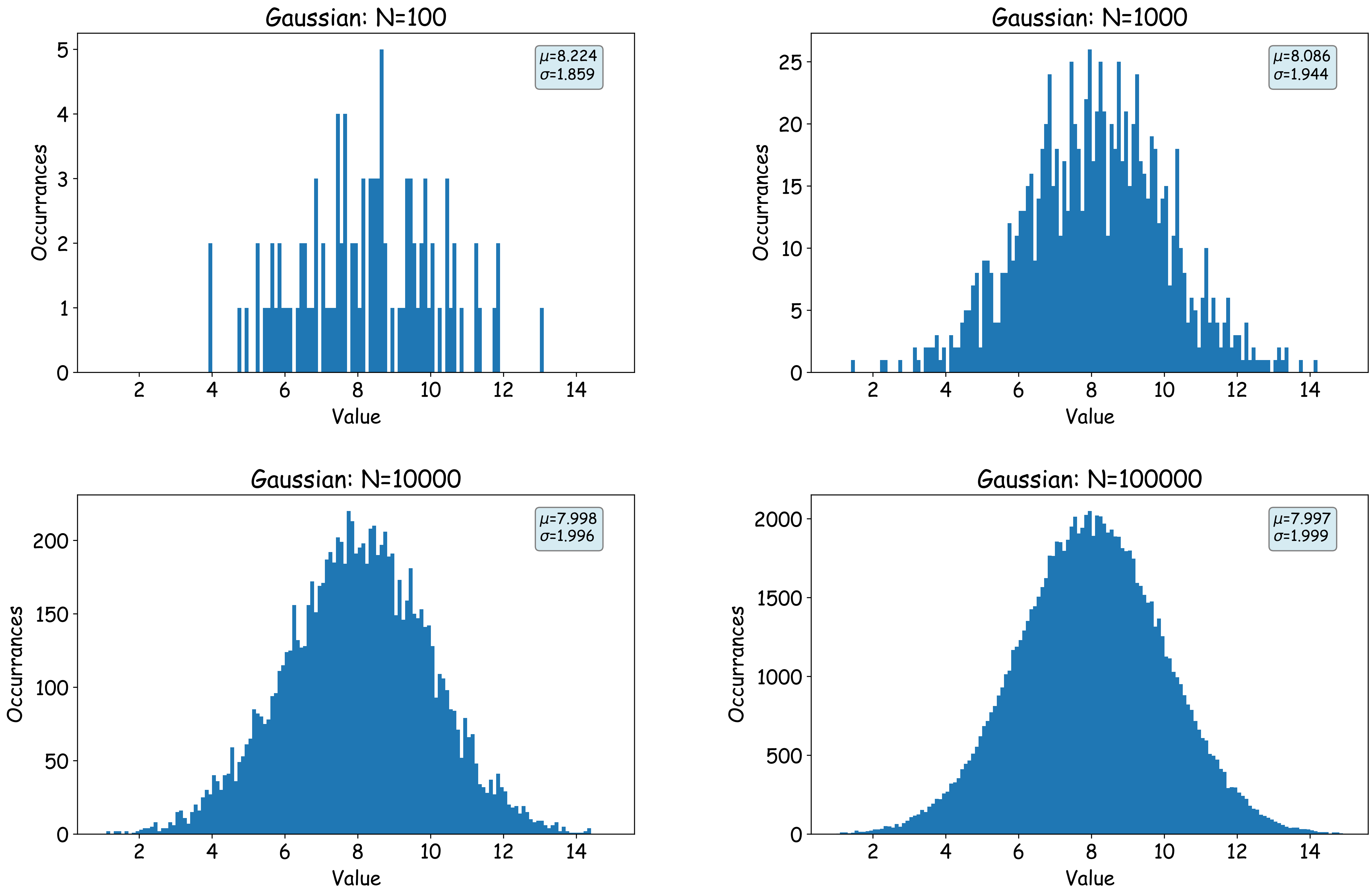
- Our best estimates of the parent population parameters (μ, σ_{parent}) are the sample mean and sample variance:

$$\mu = \frac{1}{N}\sum_i x_i \qquad \sigma_{sample} = \sqrt{\frac{\sum (x_i - \mu)^2}{N_{sample} - 1}}$$

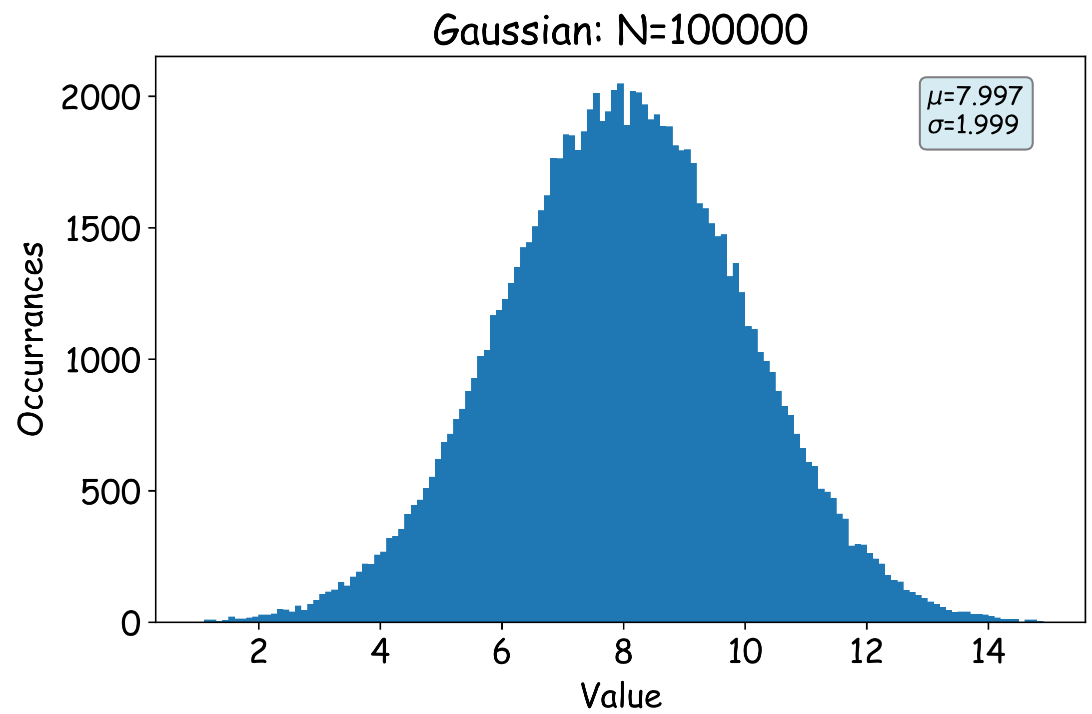
Use N-1 when mean, μ, derived from the data

The Mean and Its Error

- The more measurements of a quantity we take the more precisely we can characterise the distribution (and get closer to the parent distribution)



The Mean and Its Error

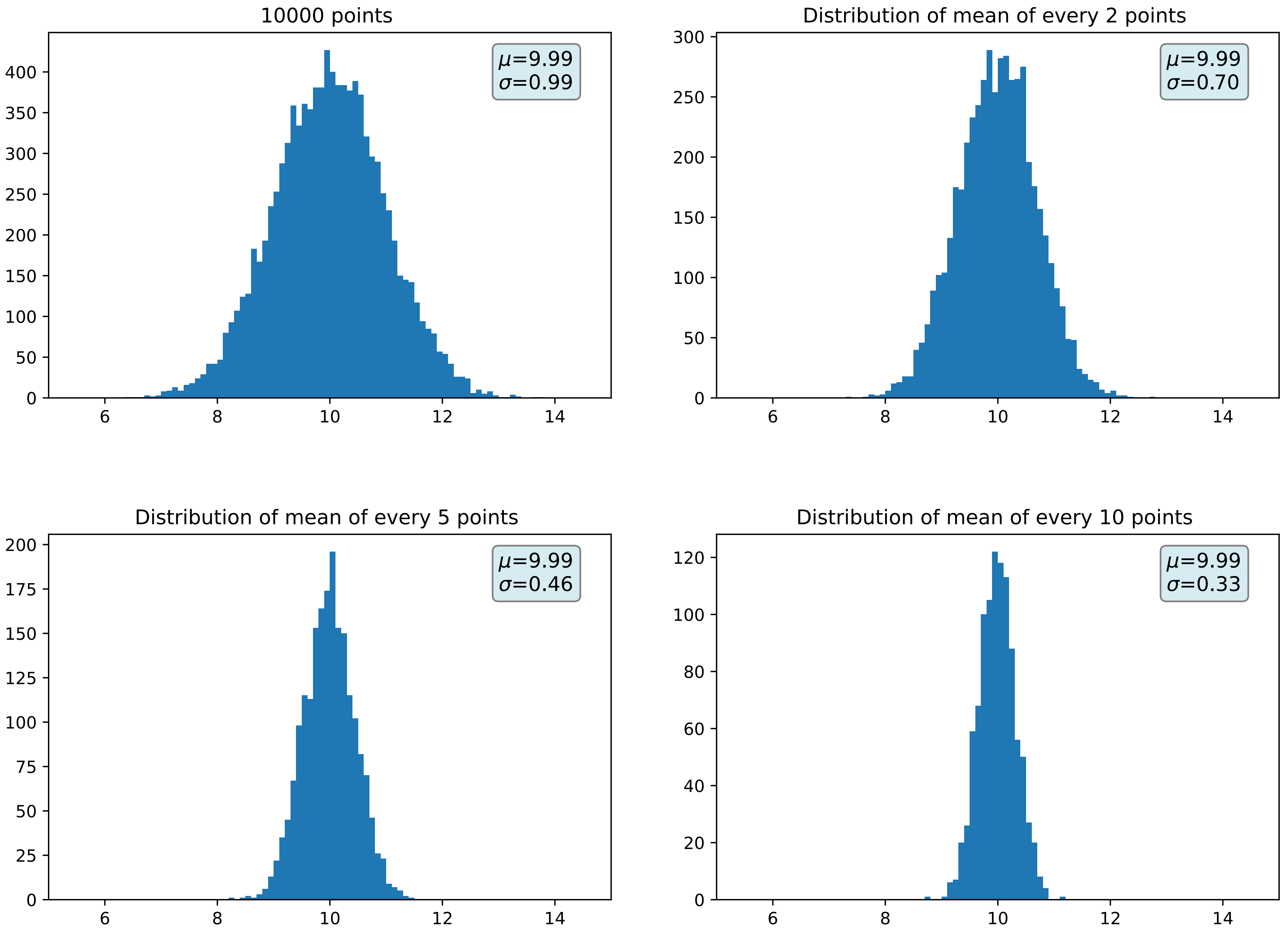


- What does this distribution tell us?
  - The probability of getting a given value in a single measurement.
- What is the uncertainty on the mean?
  - We know the mean much more accurately than to within $\pm 1\sigma$!
- Using Propagation of Errors it is possible to show: $$\sigma_\mu = \frac{\sigma}{\sqrt{N}}$$

If N data points are averaged to get a value, the error on that value is not the standard deviation (error) σ of the distribution but $\sigma/\sqrt{N}$ ("the standard error on the mean")

Averaging reduces the error

- The best way to obtain the best estimate of a value and reduce the uncertainty is to average many points.



10,000 random points from a Normal distribution (mean=10.0, std=1.0) were simulated (distribution of values shown in top left plot).

These were then grouped in 2s, 5s and 10s and the distribution of the means of those groups is shown in the other three figures.

Averaging reduces the error

- The simulation has 10,000 data points, from a Normal distribution with parent population parameters μ=10.0, σ=1.0,
- The 10,000 points were subdivided into groups of different sizes, the groups averaged, and the distribution of the means (averages) investigated.

Width of Distribution of Means

Uncertainty on overall mean

# points in each group (M)	# means (i.e. # groups) (N)	σ_{means} (calculated from histogram)	σ_{means} (expected) ($=\sigma_{parent}/\surd M$)	σ_{mean} ($=\sigma_{means}/\surd N$) (calculated)
1	10,000	0.99	1.00	0.01
2	5,000	0.70	0.71	0.01
5	2,000	0.46	0.45	0.01
10	1,000	0.33	0.32	0.01
10,0000	1	-	0.01	0.01

The Weighted Mean and Its Error

- If we have a set of measurements taken with different uncertainties (e.g. we improve the technique or apparatus part of the way through), then we can combine the data using the following formulae:

$$\mu = \frac{\sum (x_i/\sigma_i^2)}{\sum (1/\sigma_i^2)}$$

"Weighted Mean"

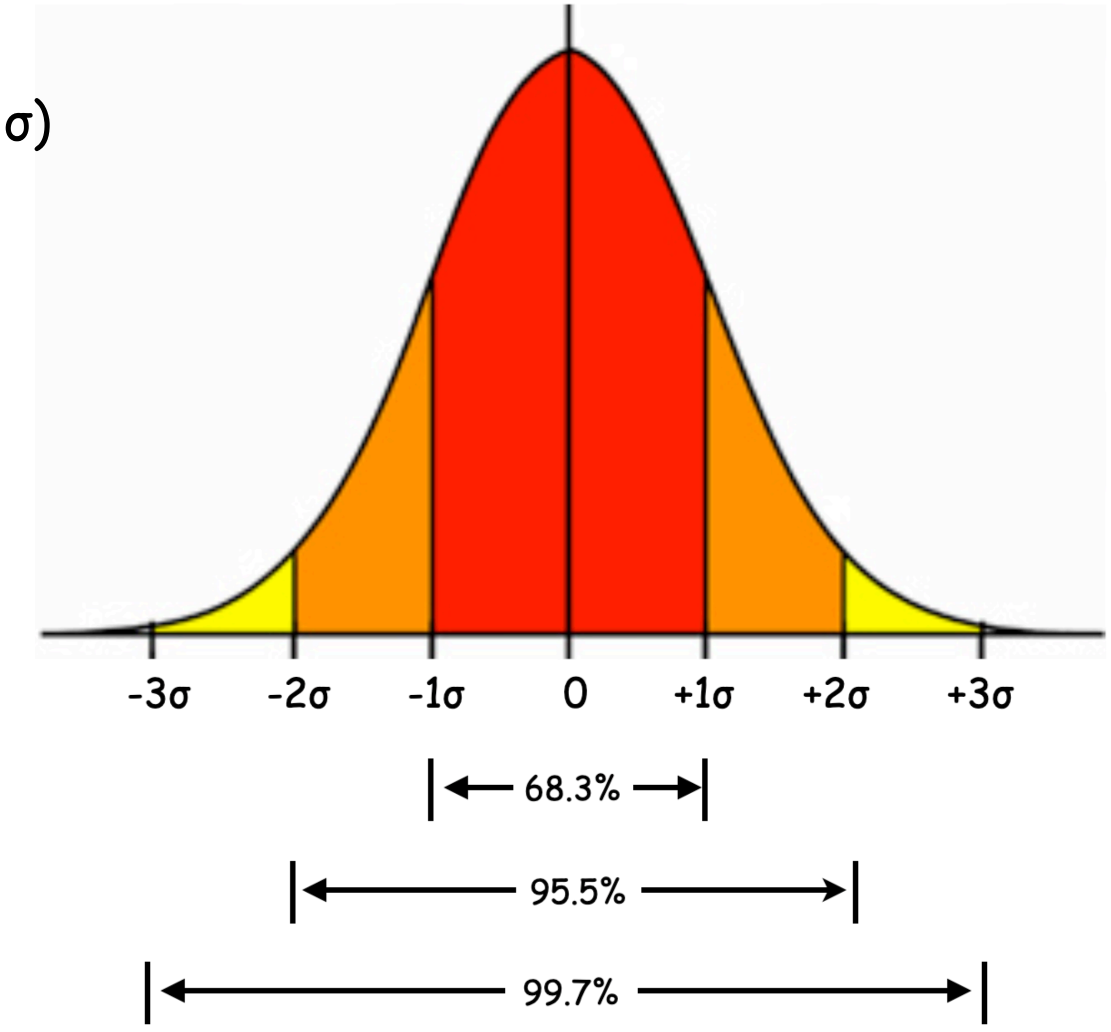
$$\sigma_\mu^2 = \frac{1}{\sum (1/\sigma_i^2)}$$

"Error on the Weighted Mean"

Comparing Experimental Results with an Accepted Value

- Calculate how many multiples of the standard error (i.e. σ) the measured value is away from the accepted value:

$$\frac{\text{measured} - \text{expected}}{\text{standard error}}$$



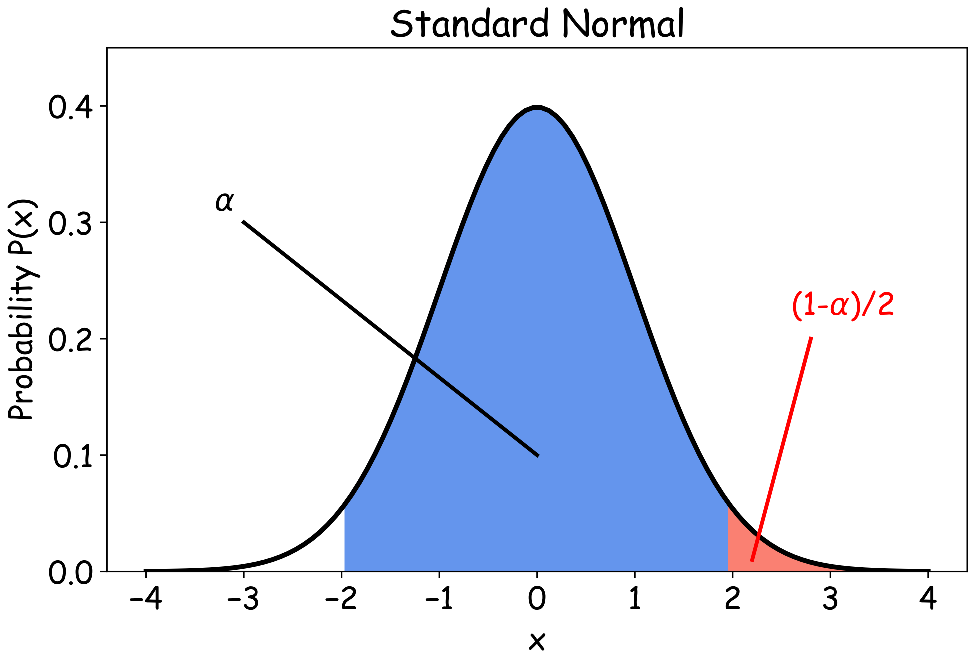
Values:

- Up to 1: excellent agreement
- 1-2: good agreement
- >3: measurements are in disagreement

- Note: caution is needed if the standard error is derived from a small number of measurements - see later slide on Student's t-distribution!

Placing Confidence Limits on a Parameter

- In the physical sciences 1σ is commonly used for the confidence region but in other areas of science 95% is common.
- If we know the mean and standard error we can then place limits at any given confidence level (α) on the value of a parameter:



  - Find the limits on the integration of the standard normal distribution to give the appropriate confidence level:

$$\int_{-x}^{+x} P_{DF}(x)\,dx = \alpha$$

  - We can use the inverse survival function to find the value of x which gives

$$\int_{x}^{+\infty} P_{DF}(x)\,dx = \frac{1-\alpha}{2}$$

```
alpha=0.95
x=norm.isf((1-alpha)/2)
print(f"95% confidence limit: {x:.3f}”)
95% confidence limit: 1.960
```

- Now the 95% limit on the parameter can be quoted as:

$$\bar{x} \pm 1.960 \times \frac{\sigma_{sample}}{\sqrt{N}}$$

- Note: caution is needed if the standard error is derived from a small number of measurements - see later slide on Student’s t-distribution!

Student's t-Distribution

- We previously checked agreement between a measurement and an expected value using:

$$\frac{\text{measured} - \text{expected}}{\text{standard error}}$$

- If the standard error is:
  - known from parent distribution then the above number is called the z-statistic
  - derived from the data sample itself from N data points then the above number is called the t-statistic

$$z = \frac{\bar{x} - \text{expected}}{\sigma_{parent}} \qquad t = \frac{\bar{x} - \text{expected}}{\frac{\sigma_{sample}}{\sqrt{N}}}$$

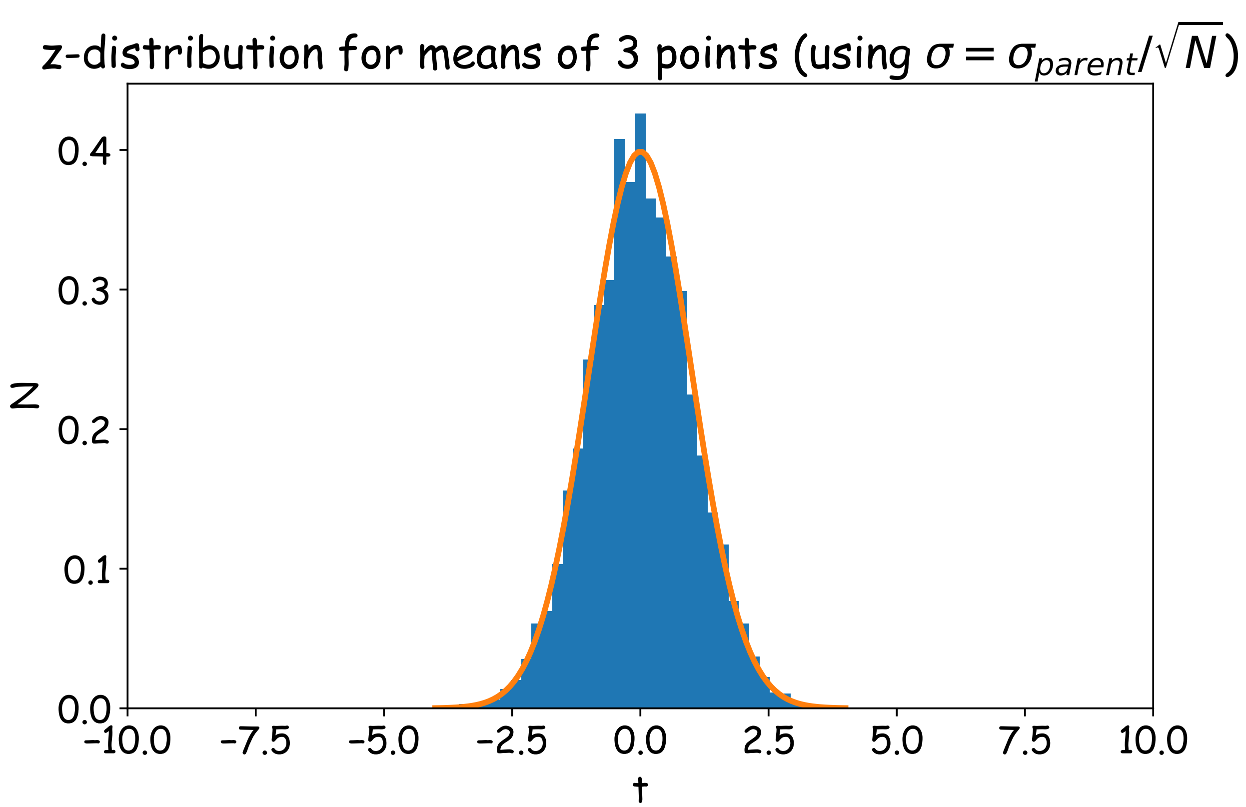
- The z-statistic is expected to be normally distributed.

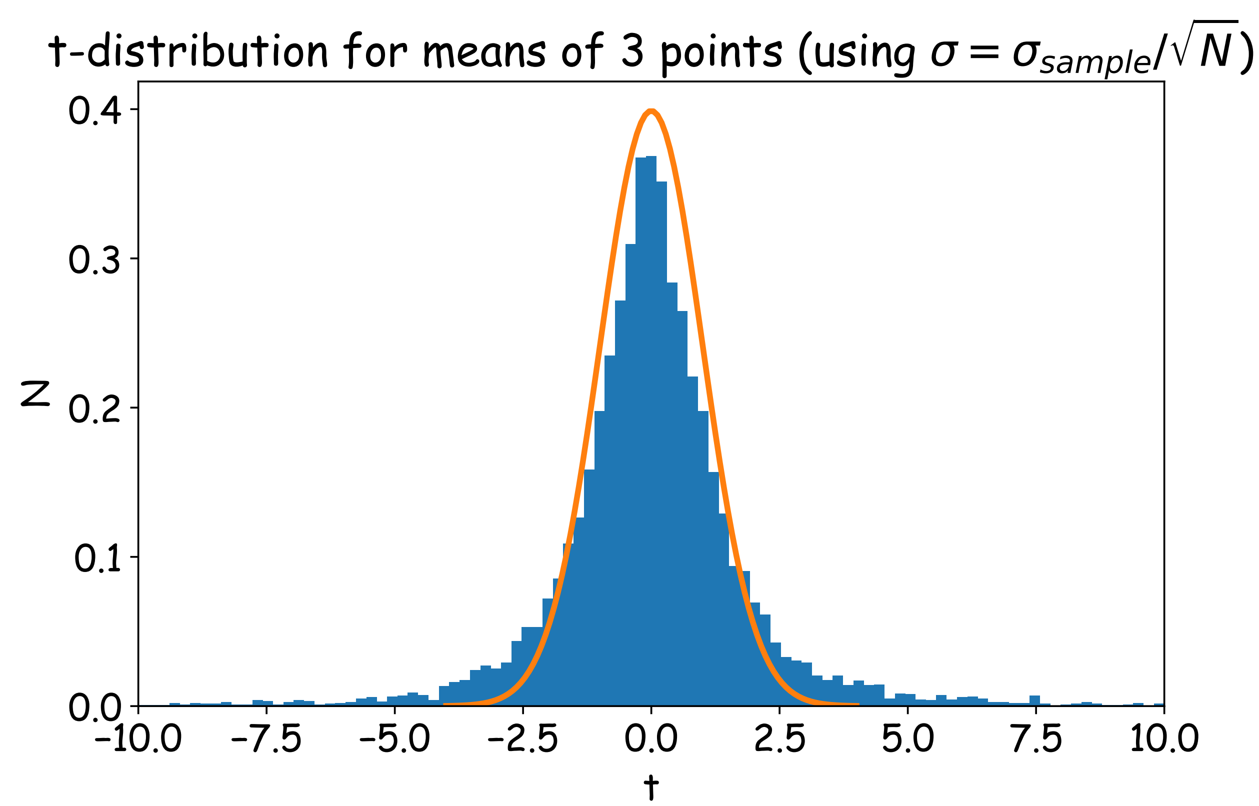
- However, if the mean and the error are derived from a small number of data points then they can be both be subject to random fluctuations which results in the t-statistic having larger deviations from the mean than expected from the normal distribution.

Student's t-Distribution

$$z = \frac{\bar{x} - \text{expected}}{\frac{\sigma_{parent}}{\sqrt{N}}} \qquad t = \frac{\bar{x} - \text{expected}}{\frac{\sigma_{sample}}{\sqrt{N}}}$$

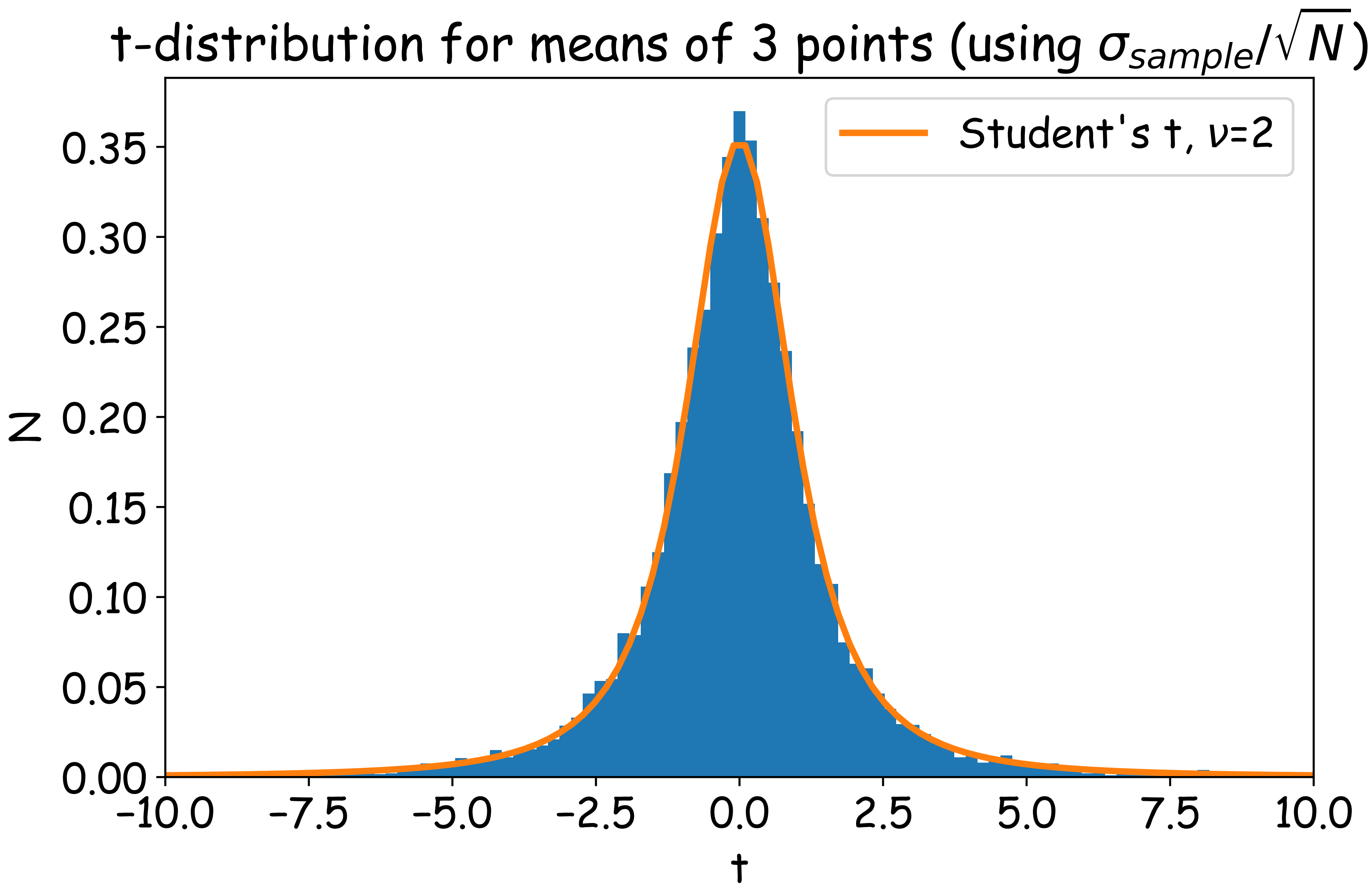
- Simulation:
  - 10,000 points from a Standard Normal
  - group into threes and calculate mean and standard deviation of every group
  - produce histograms of z-statistic and t-statistic and compare to Normal distribution





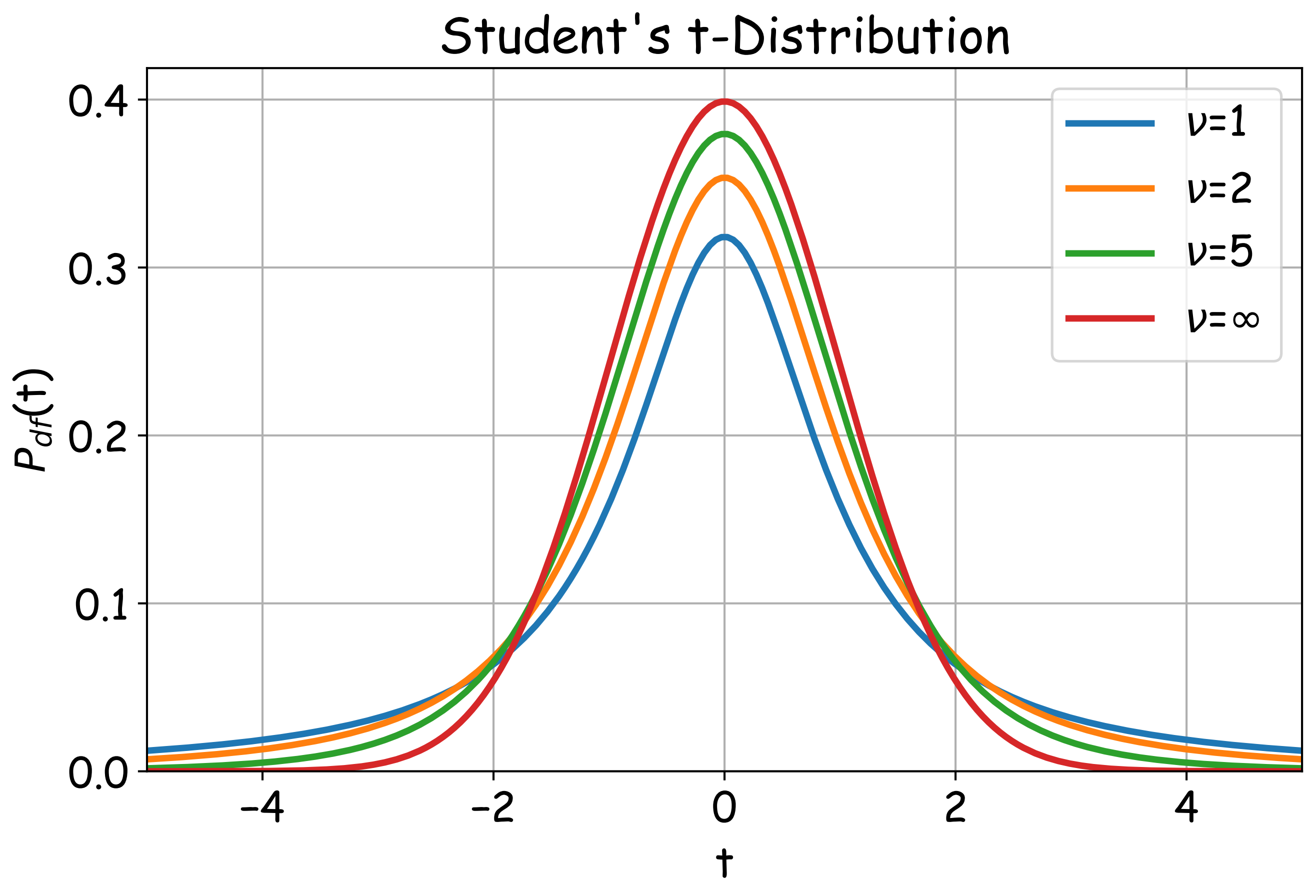
Student's t-Distribution

- The actual distribution of the t-statistic follows a distribution known as Student's t-distribution
- "Student" was the pseudonym of W. Gosset who was required to publish his work on quality control of ingredients anonymously while working at Guinness Breweries in Dublin in 1906 - wikipedia).



Student's t-Distribution

- Student's t-distribution has much wider tails than the normal distribution for small numbers of the degrees of freedom ($\nu = N - 1$)
  - and approaches the normal distribution when the number of degrees of freedom increases.
- Due to the tails the 95% or 99% confidence limits are much larger than for a Gaussian!



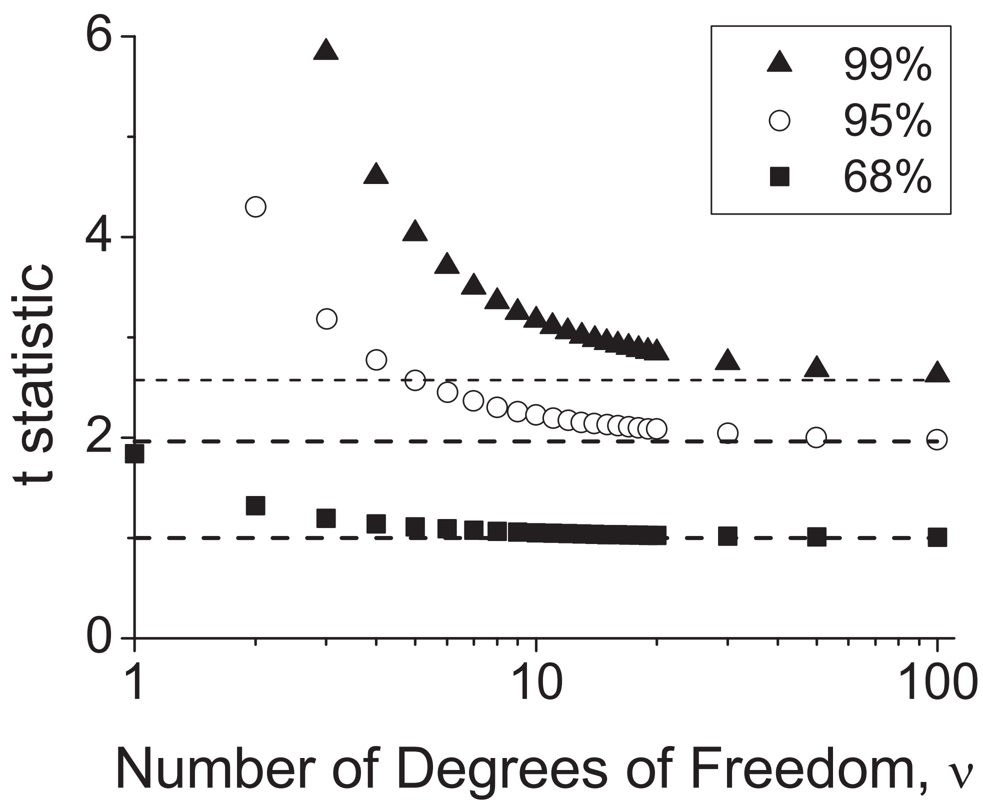
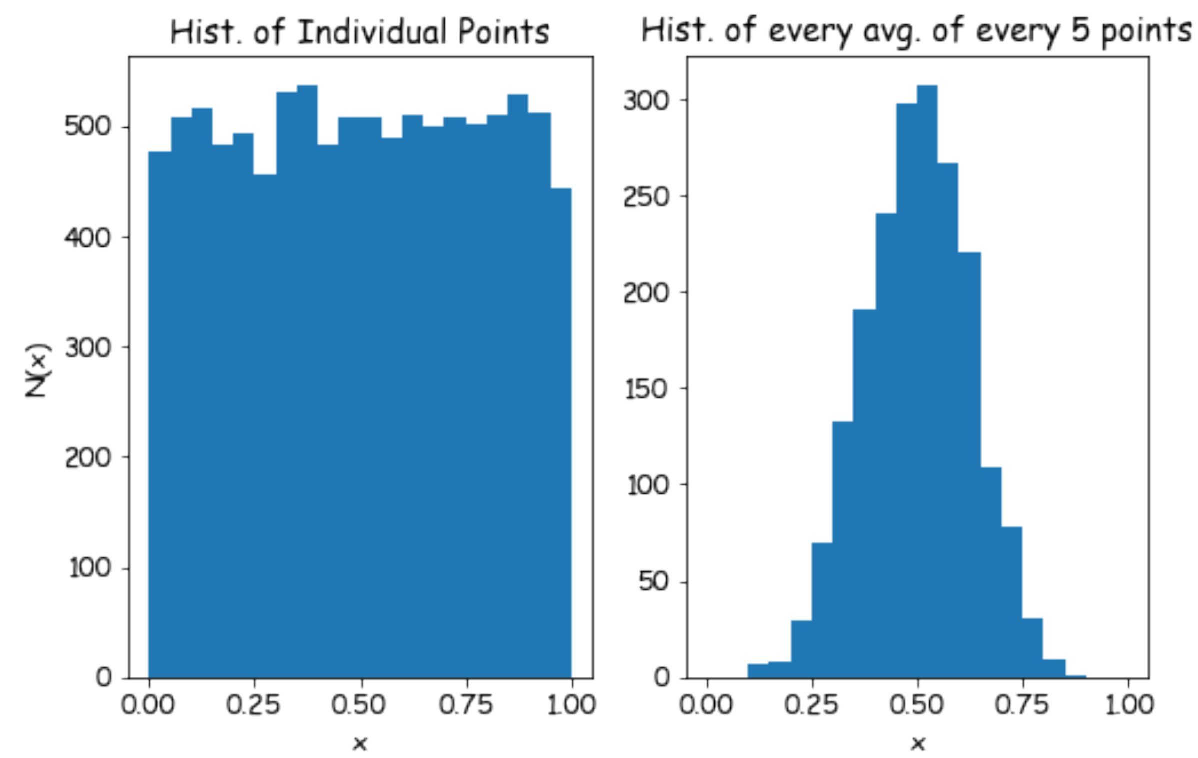


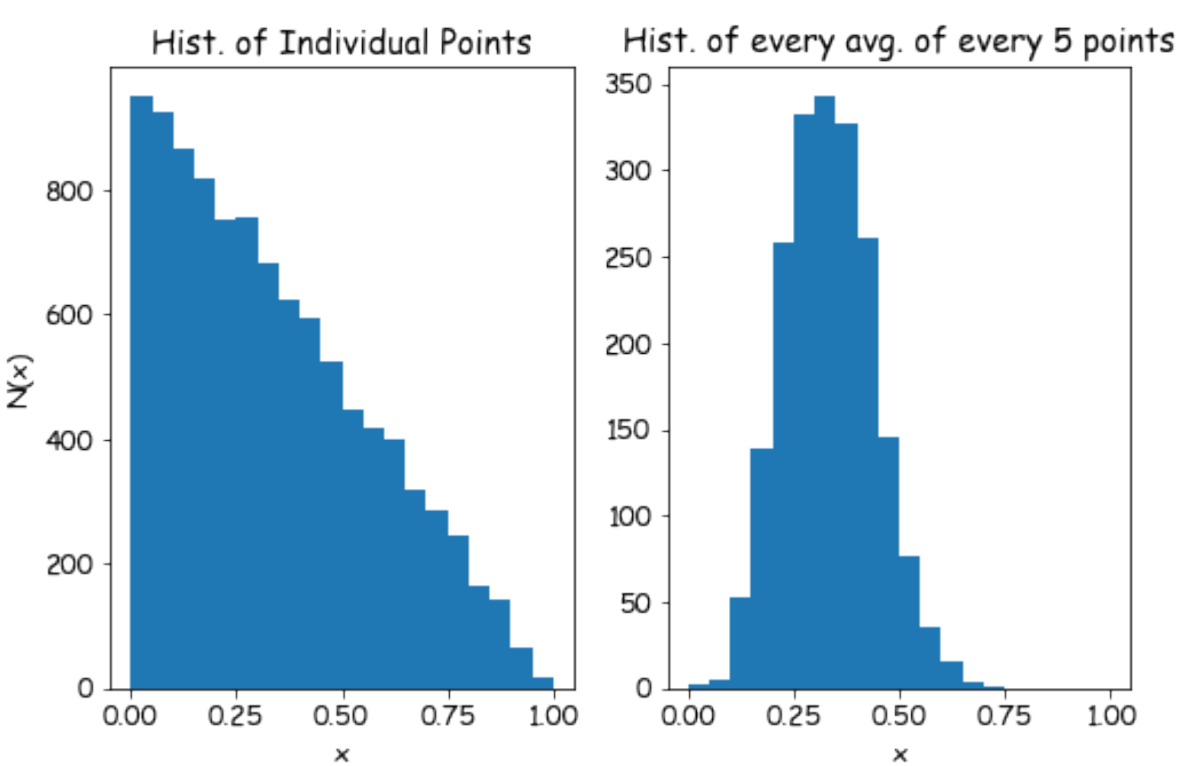
figure 8.11 from Hughes & Hase book

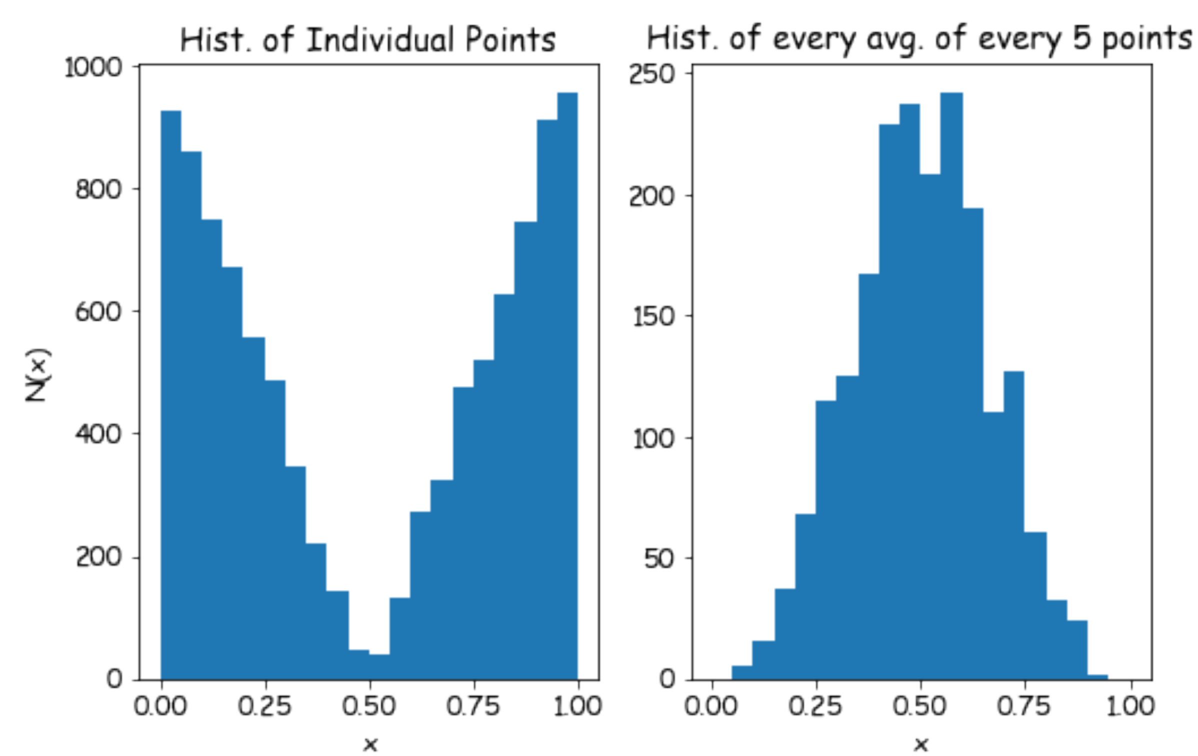
(dashed lines are expected from normal distribution for ~1σ, ~2σ and ~3σ)

Central Limit Theorem

- Why is the normal distribution so prevalent?
  - Ans: The Central Limit Theorem
    - Irrespective of the parent distribution of some variable, the distribution of the mean of that variable tends towards a normal distribution with the same mean, as the number of samples becomes large (not many needed for many 'reasonable functions'!)

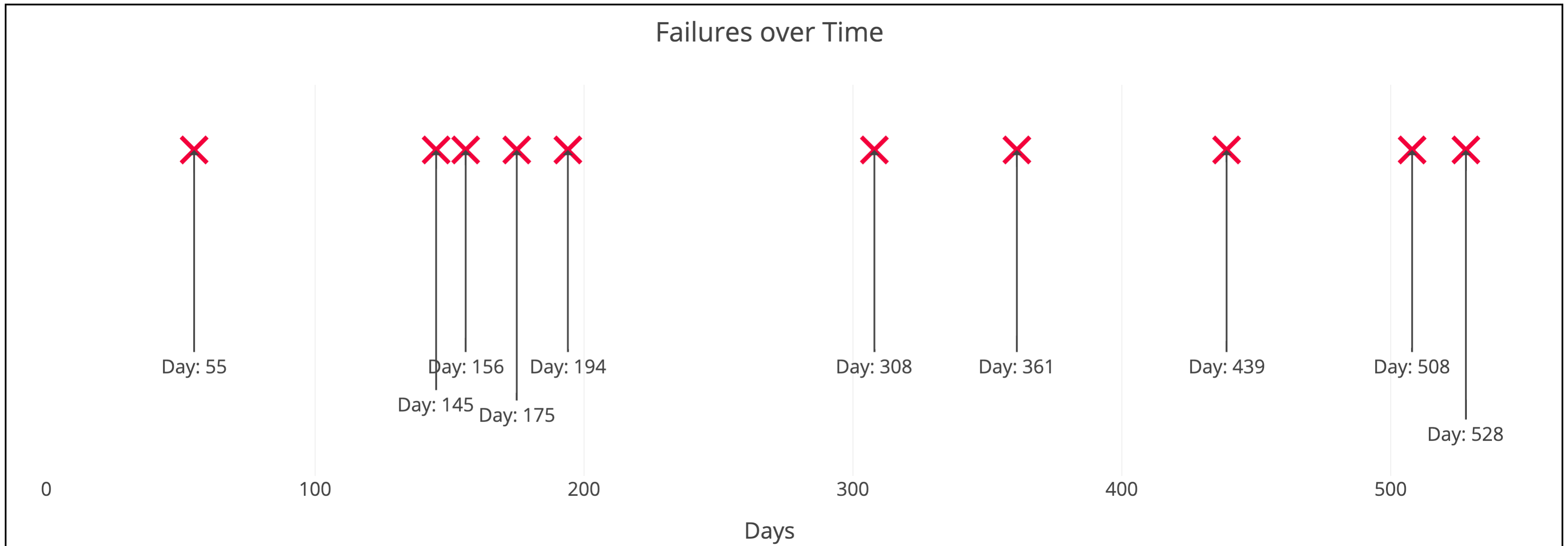






Counting Experiments (Poisson Distr.)

- When one counts the number of random events in an interval (time, area, volume, etc) and repeats the experiment under identical conditions then one does not always get the same result.



https://towardsdatascience.com/the-poisson-distribution-and-poisson-process-explained-4e2cb17d459

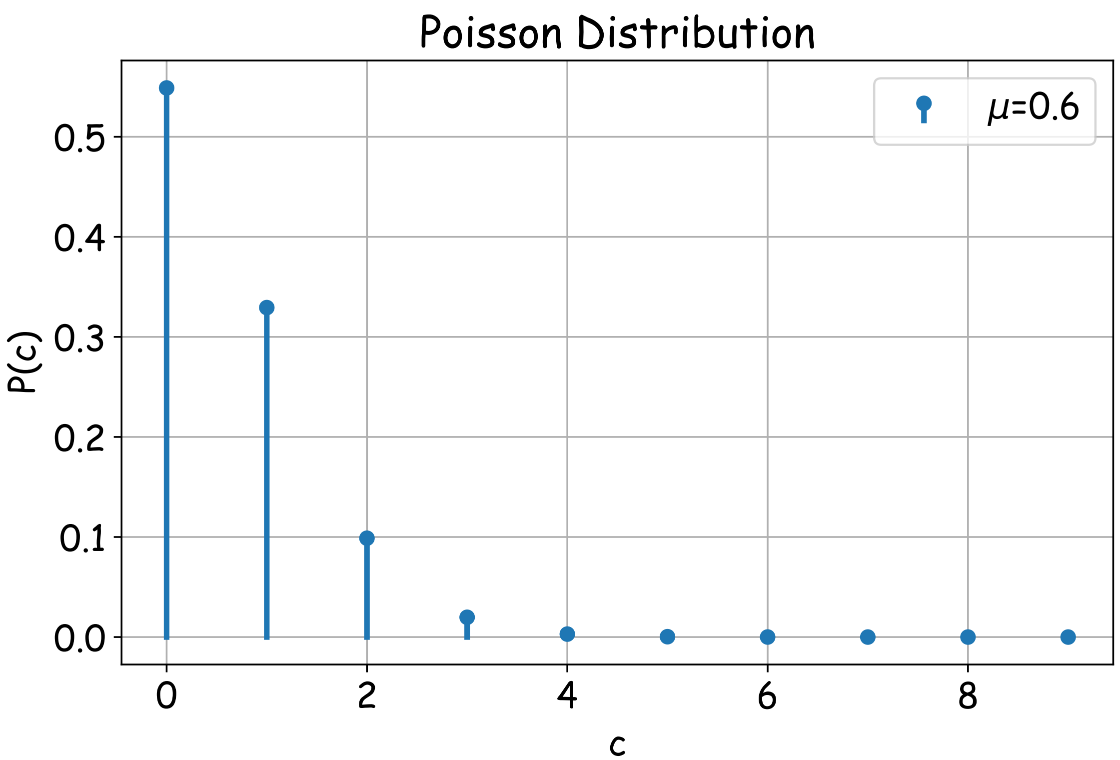
- The distribution of random counted events follows a Poisson Distribution:

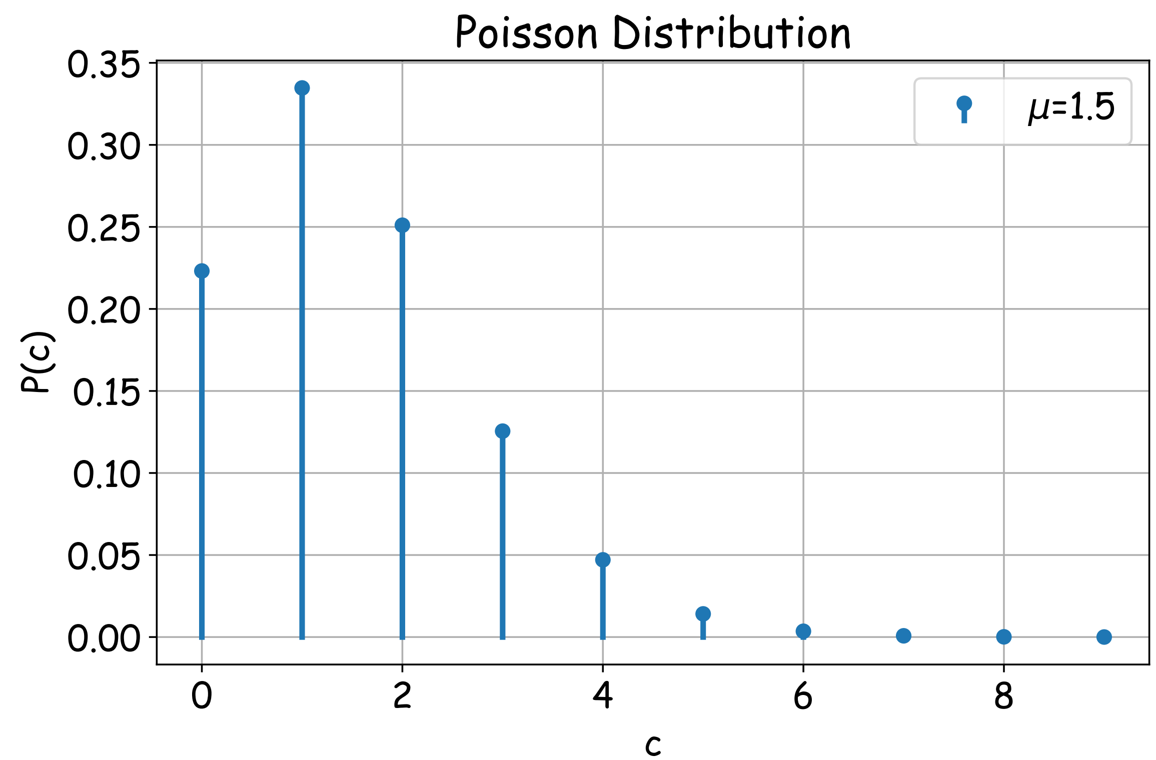
$$P(n) = \frac{\mu^n e^{-\mu}}{n!}$$

where,

- $P(n)$ is the probability of obtaining n events in a given interval
- μ is the mean of the distribution.
- Note: the standard deviation has value $\sqrt{\mu}$

Counting Experiments (Poisson Distr.)





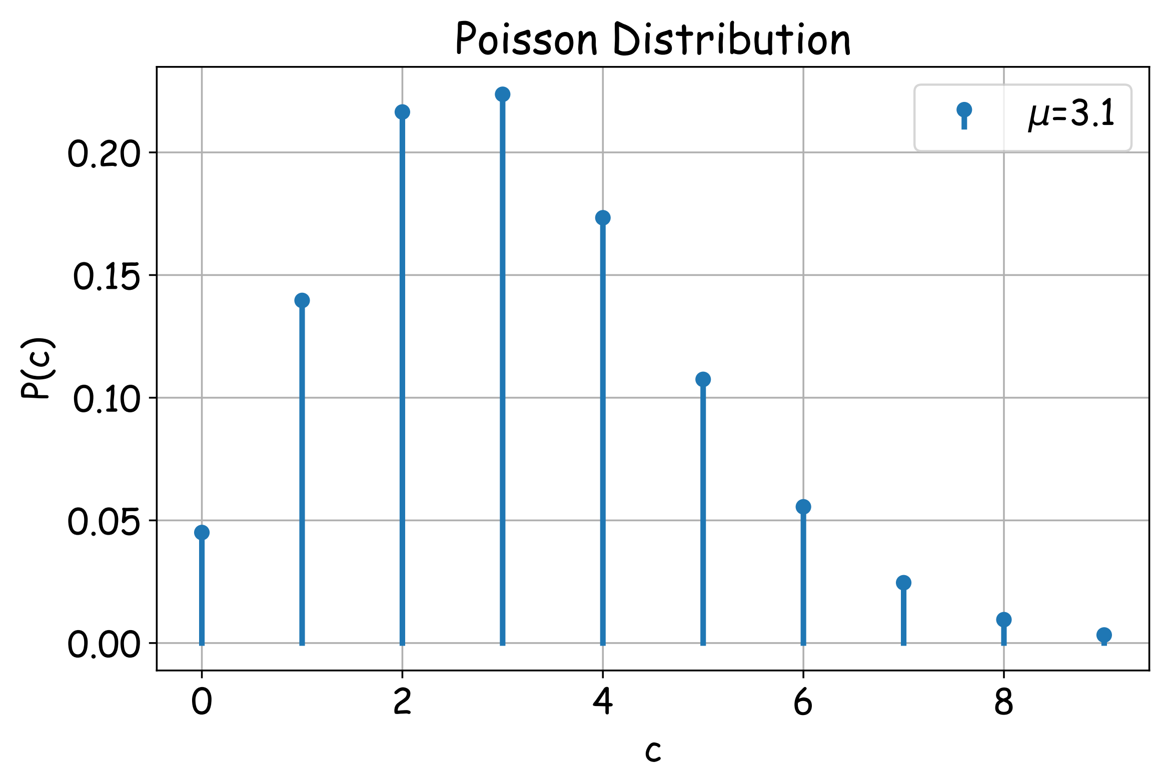
Asymmetric!

μ=0.6

$\sigma=\sqrt{0.6}=0.775$

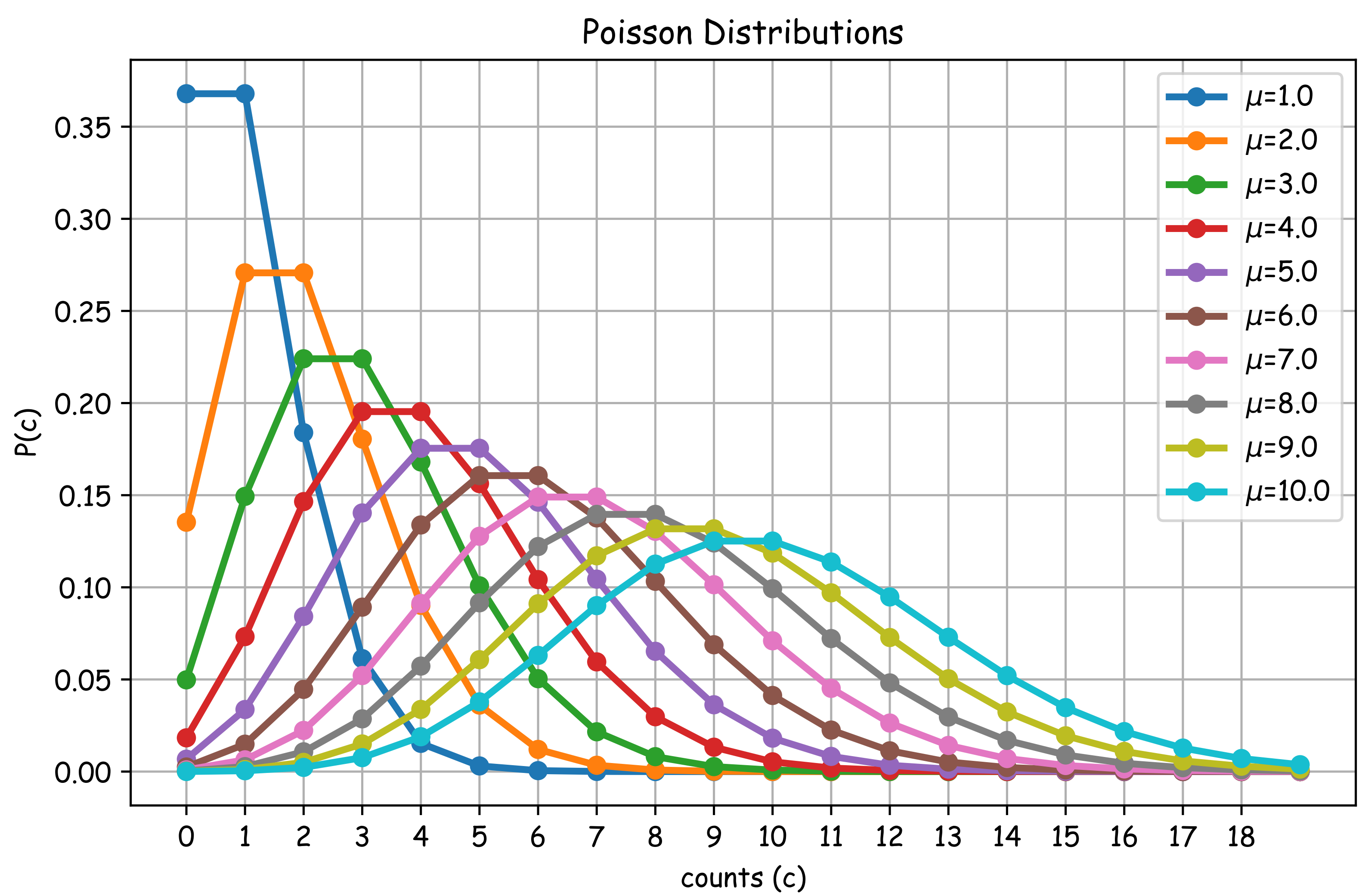
It does not make sense to quote: 0.600 ± 0.775

We need asymmetric error bars!



Counting Experiments (Poisson Distr.)

- For small means the distribution is asymmetric.
- For means ≳10 the distribution is symmetric and is approximately described by a Gaussian distribution of mean μ and standard deviation $\sqrt{\mu}$.



- For dealing with errors (especially propagation) in counting experiments we generally want enough counts for the Poisson distribution to be in Gaussian regime.

Binomial Distribution

- The Binomial Distribution describes the probability of observing k successes out of n tries where the probability of success p.

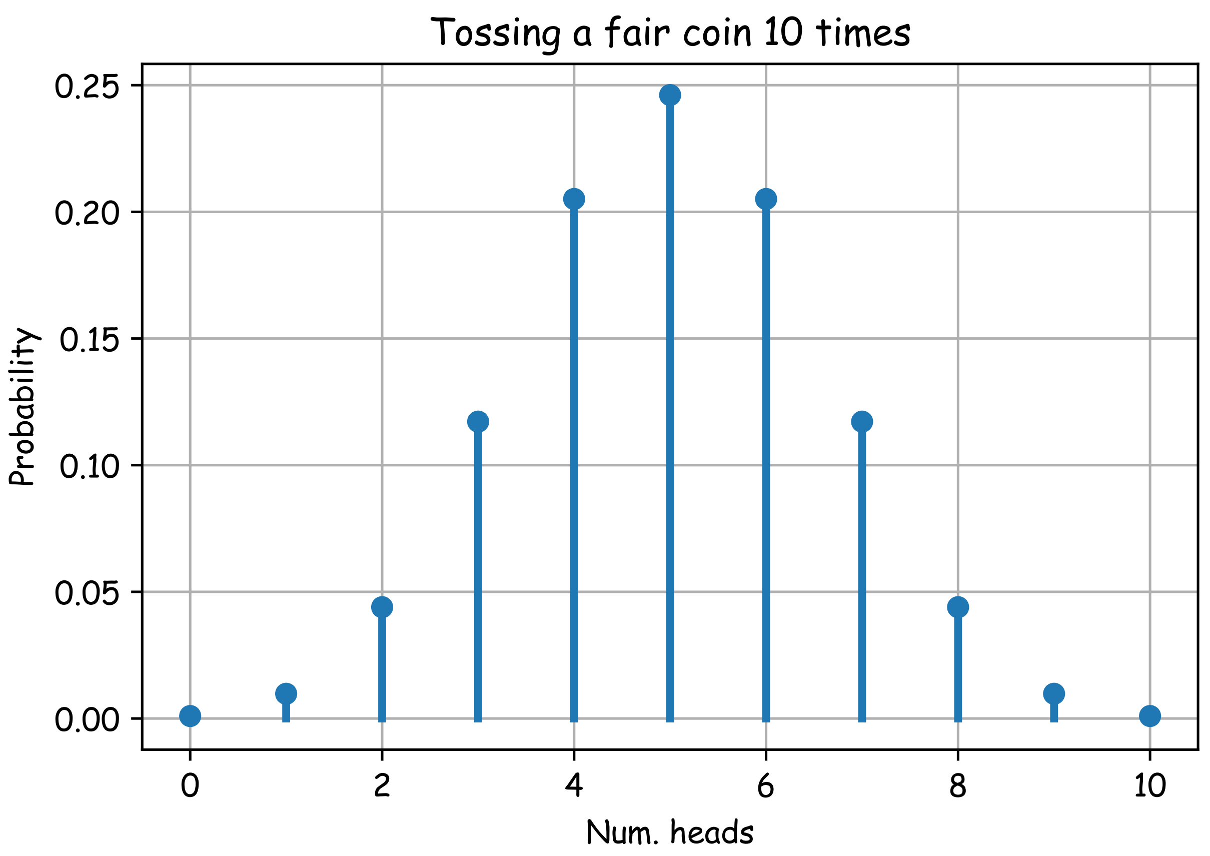
- The distribution is described by the equation:

$$P_B(k,n,p) = \binom{n}{k} p^k\, q^{n-k} = \frac{n!}{k!(n-k)!} p^k\, (1-p)^{n-k}$$

$$\mu = \sum x P_B(x) \qquad \sigma^2 = \sum (x-\mu)^2 P_B(x)$$

e.g. tossing a fair coin 10 times - what is the probability of getting 0, 1, 2, ... 10 heads?

`scipy.stats.binom`



- the Normal distribution may be used to approximate the binomial distribution when np>>1.

Conclusions

- A crucial part of experimental science is proper evaluation and estimation of uncertainties.
- Measurements generally have a spread and are characterised by a distribution:
  - Normal/Gaussian is most common - due to the Central Limit Theorem
    - characterised by two parameters: mean (μ) and standard deviation (σ)
  - Poisson for counting experiments
    - characterised by a single parameter: μ
    - if counts are large enough then the Poisson distribution is in the Gaussian regime (with mean μ and standard deviation $\sigma=\sqrt{\mu}$)
- Where repeated measurements are possible:
  - The best estimate of a parameter is the mean
  - The (sample) standard deviation characterises the spread in individual measurements
  - The error on the estimated parameter is the error on the mean ($\sigma_{sample}/\sqrt{N}$)
- We can use the probability distribution to estimate the uncertainty on a parameter at a given confidence level, and to test the agreement with an expected/predicted value.