Class 1: Probability & Statistics

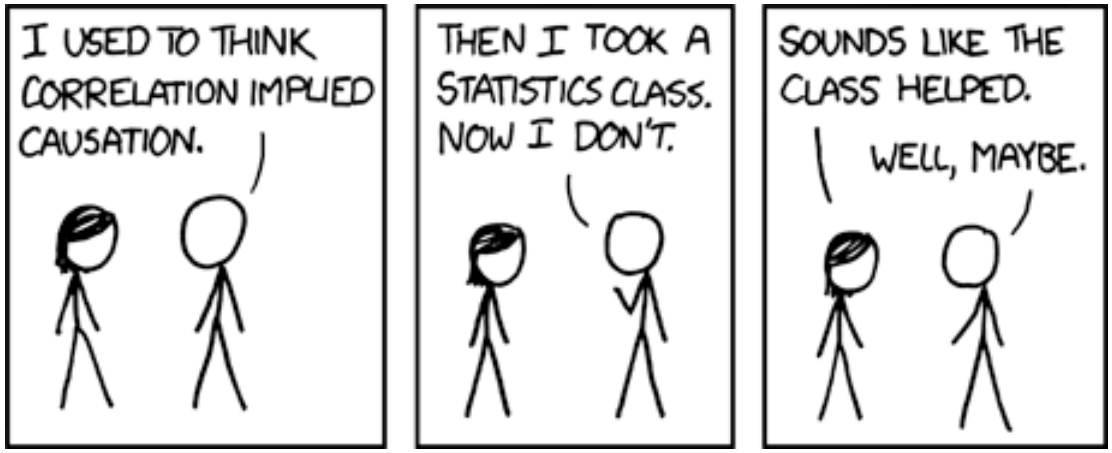
In this class we will review how statistics are used to summarize data, special probability distributions, their use in simple applications using Frequentist and Bayesian methods, and Monte Carlo techniques

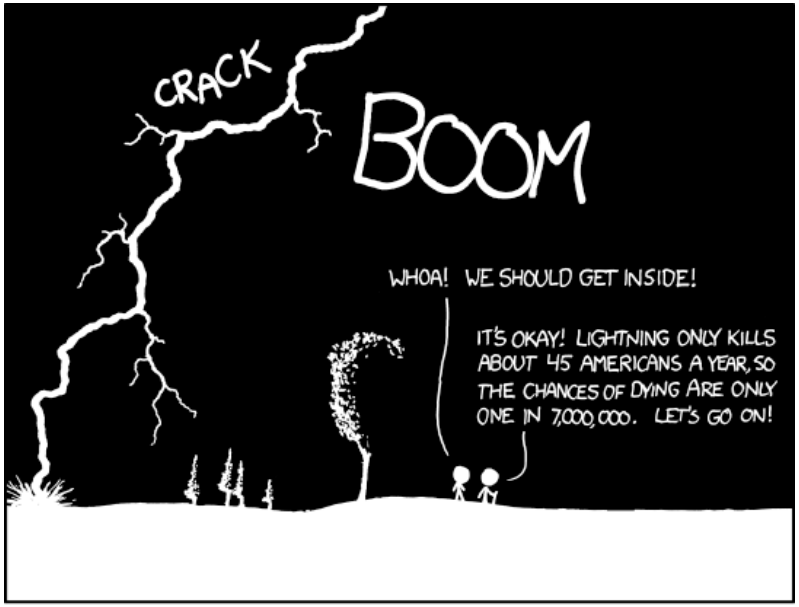
Class 1: Probability & Statistics

At the end of this class you should be able to …

- … determine summary statistics for datasets and their errors
- … optimally combine data
- ... apply probability distributions for Gaussian, Binomial and Poisson statistics
- … compare the Frequentist and Bayesian frameworks for statistical analysis
- … solve statistical problems using Monte Carlo techniques

Class 1: Probability & Statistics



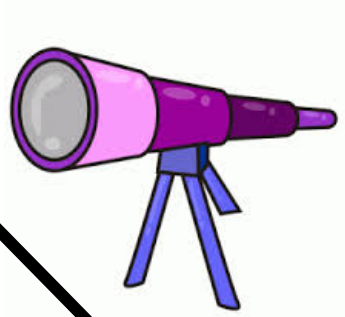


THE ANNUAL DEATH RATE AMONG PEOPLE WHO KNOW THAT STATISTIC IS ONE IN SIX.

Credit: xkcd.com

The process of science

Obtain measurements



Design a question

Analyze data

Conclude (test hypothesis, change probabilities)

The point of statistics

"If your experiment needs statistics, you ought to have done a better experiment" (E.Rutherford)

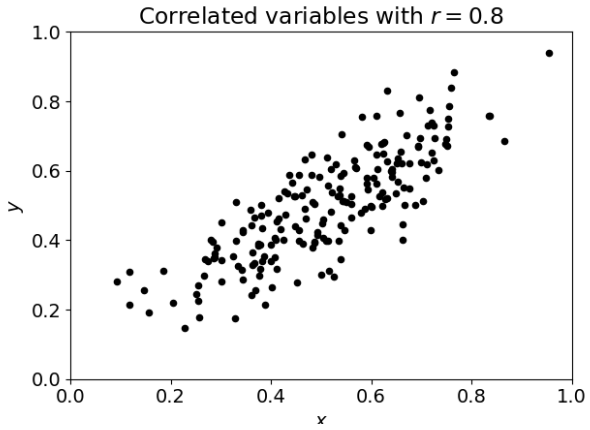
"A body of methods for making wise decisions in the face of uncertainty" (W.Wallis)

- Statistics allows us to formulae the logic of **what** we are doing and **why**. It allows us to make **precise statements**.
- Statistics allows us to quantify the **uncertainty** in any measurement (which should always be stated)
- Statistics allows us to avoid pitfalls such as **confirmation bias** (distortion of conclusions by preconceived beliefs)

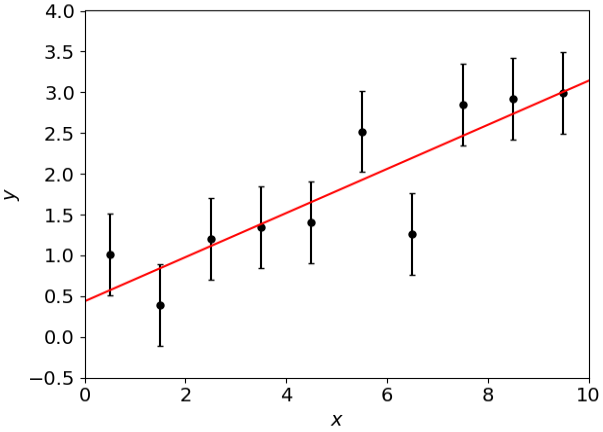
Common uses of statistics

- **Measuring a quantity ("parameter estimation"):** given some data, what is our best estimate of a particular parameter? What is the uncertainty in our estimate?

- **Searching for correlations**: are two variables we have measured correlated with each other, implying a possible physical connection?



- **Testing a model ("hypothesis testing")**: given some data and one or more models are our data consistent with the models? Which model best describes our data?



Summary statistics and their errors

- A **statistic** is a quantity which summarizes our data

Image credit: pythonstatistics.net

Summary statistics and their errors

- A **statistic** is a quantity which summarizes our data
- I have a sample of N independent estimates x_i of some quantity, how can I summarize them?
- The **mean** (typical value): $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
- The **median** (middle value when ranked)
- The **standard deviation** σ (spread) or **variance**:

$$\mathrm{Var}(x) = \sigma^2(x) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

- [Small print: Watch out for factor of $N-1$! (see below)]

Summary statistics and their errors

- We can quote an **error** in each of these statistics:
- Error in the mean is **standard deviation divided by** $\sqrt{N}$ (as I increase the sample size, the error in the mean improves)

$$\text{Error in mean} = \frac{\sigma}{\sqrt{N}}$$

- Error in the median $= 1.25\ \frac{\sigma}{\sqrt{N}}$
- Error in the variance $= \sigma^2 \sqrt{\frac{2}{N-1}}$
- [Small print: the error in the mean holds for all probability distributions. The other two relations assume a Gaussian distribution.]

Estimators and bias

- These formulae are a good example of **estimators**, *combinations of data which measure underlying quantities*

- E.g., the estimator $\hat{V} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$ measures the underlying variance V [notice "hat" notation meaning "estimate of"]

- If an estimator is **unbiased**, then it recovers the true value **on average** over many realisations of the data, $\langle \hat{V} \rangle = V$ [notice notation $\langle \ldots \rangle$ meaning "average over many experiments"]

- [Small print: we can show that the $\frac{1}{N-1}$ factor in $\hat{V}$ is needed to ensure it is unbiased (because $\bar{x}$ is estimated from the data itself).]

Optimal combination of data

- A common statistical task is to **combine** different input data into a single measurement
- In this process we may give inputs different **weights**

Optimal combination of data

- Suppose we have N independent estimates x_i of some quantity y, which have varying errors σ_i. What is our best combined estimate of y?
- A simple average, $\hat{y} = \frac{1}{N}\sum_{i=1}^{N} x_i$?
- This is not the optimal combination, because we want to give **more weight to the more precise estimates**. Let's weight each estimate by w_i:

$$\hat{y} = \frac{\sum_{i=1}^{N} w_i\, x_i}{\sum_{i=1}^{N} w_i}$$

- [Small print: this estimate is unbiased, since $\langle \hat{y} \rangle = \frac{\sum_i w_i \langle x \rangle}{\sum_i w_i} = \langle x \rangle = y$]

Optimal combination of data

- The weights which minimize the combined error are **inverse-variance weights** $w_i = 1/{\sigma_i}^2$

$$\hat{y} = \frac{\sum_{i=1}^{N} x_i/{\sigma_i}^2}{\sum_{i=1}^{N} 1/{\sigma_i}^2}$$

- In this case, the variance in the combined estimate is:

$$\frac{1}{\mathrm{Var}(\hat{y})} = \sum_{i=1}^{N} \frac{1}{{\sigma_i}^2}$$

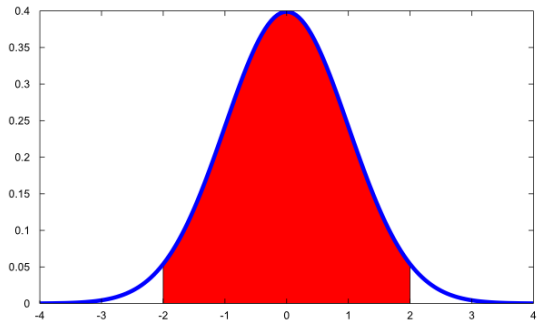
- [Small print: this approach is only helpful if the errors in the data are dominated by statistical, not systematic errors]

Worked examples

- We have $N = 10$ measurements of a variable $x_i = (7.6, 5.8, 8.0, 6.9, 7.2, 7.5, 6.4, 8.1, 6.3, 7.0)$. Estimate the mean, variance and median of this dataset. What are the errors in your estimates?
- We have $N = 5$ measurements of a quantity: $(7.4 \pm 2.0, 6.5 \pm 1.1, 4.3 \pm 1.7, 5.5 \pm 0.8, 6.0 \pm 2.5)$. What is the optimal estimate of this quantity and the error in that estimate?
- A further measurement 3.0 ± 0.2 is added. How should our estimate change?
- How can we check the reliability of the initial 5 measurements?

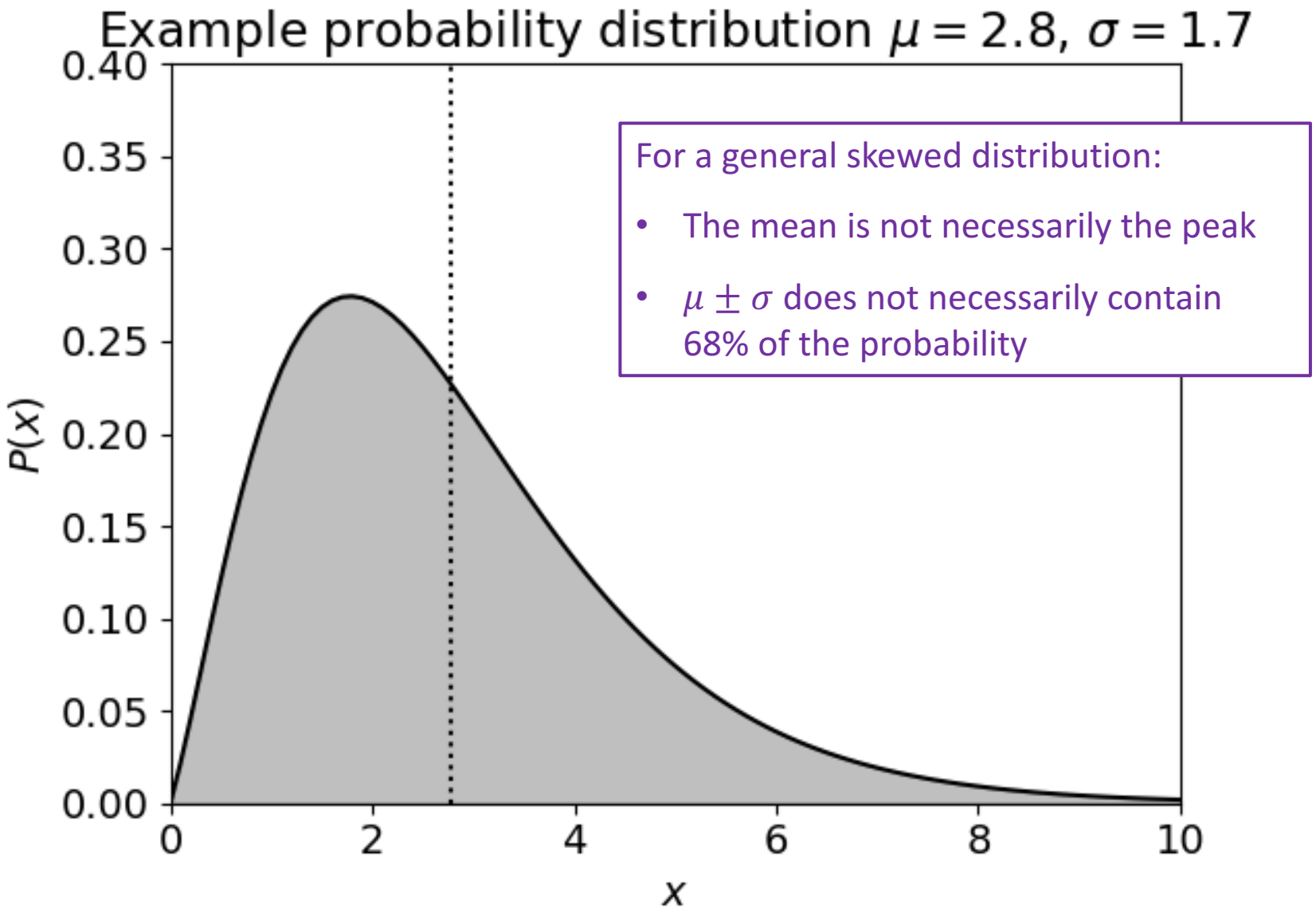
Probability distributions

- A **probability distribution**, $P(x)$, is a function which assigns a probability for each particular value (or range of values) of a continuous variable x
- Must be **normalized**: $\int_{-\infty}^{\infty} P(x)\,dx = 1$
- Probability in range $[x_1, x_2] = \int_{x_1}^{x_2} P(x)\,dx$



- A probability distribution may be quantified by its …
- **Mean** $\mu = \bar{x} = \langle x \rangle = \int_{-\infty}^{\infty} x\,P(x)\,dx$
- **Variance** $\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2\,P(x)\,dx = \langle x^2 \rangle - \langle x \rangle^2$

Probability distributions

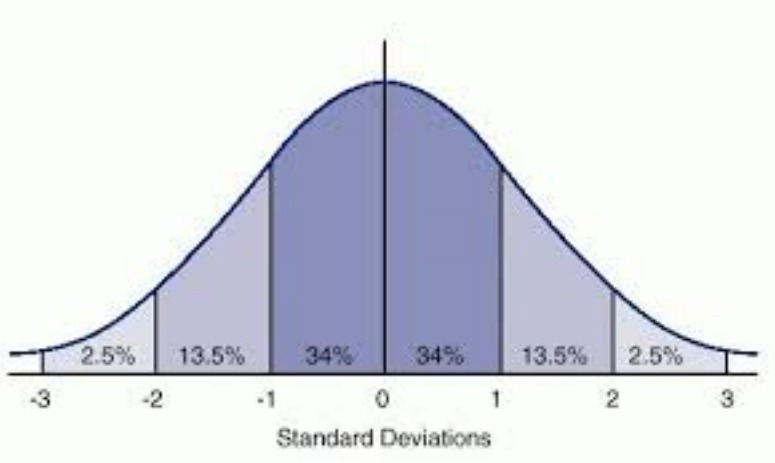


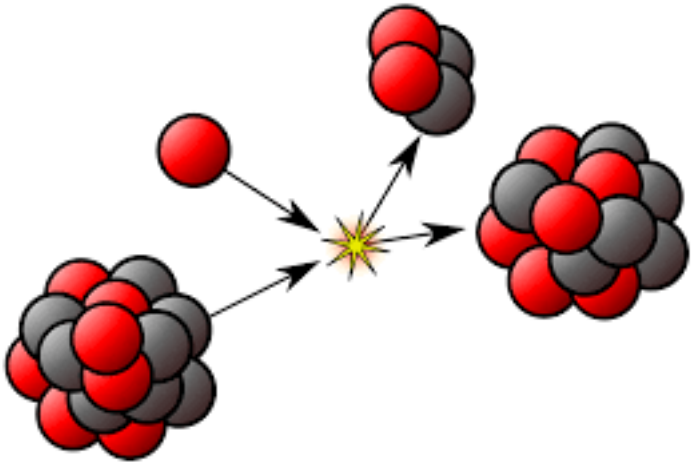
For a general skewed distribution:

- The mean is not necessarily the peak
- $\mu \pm \sigma$ does not necessarily contain 68% of the probability

Probability distributions

- Certain types of variables have known distributions:
- **Binomial** distribution
- **Poisson** distribution
- **Gaussian** or **Normal** distribution





The Binomial distribution

- Applies in problems where there is a random process with **two possible outcomes** with probabilities p and $1 - p$

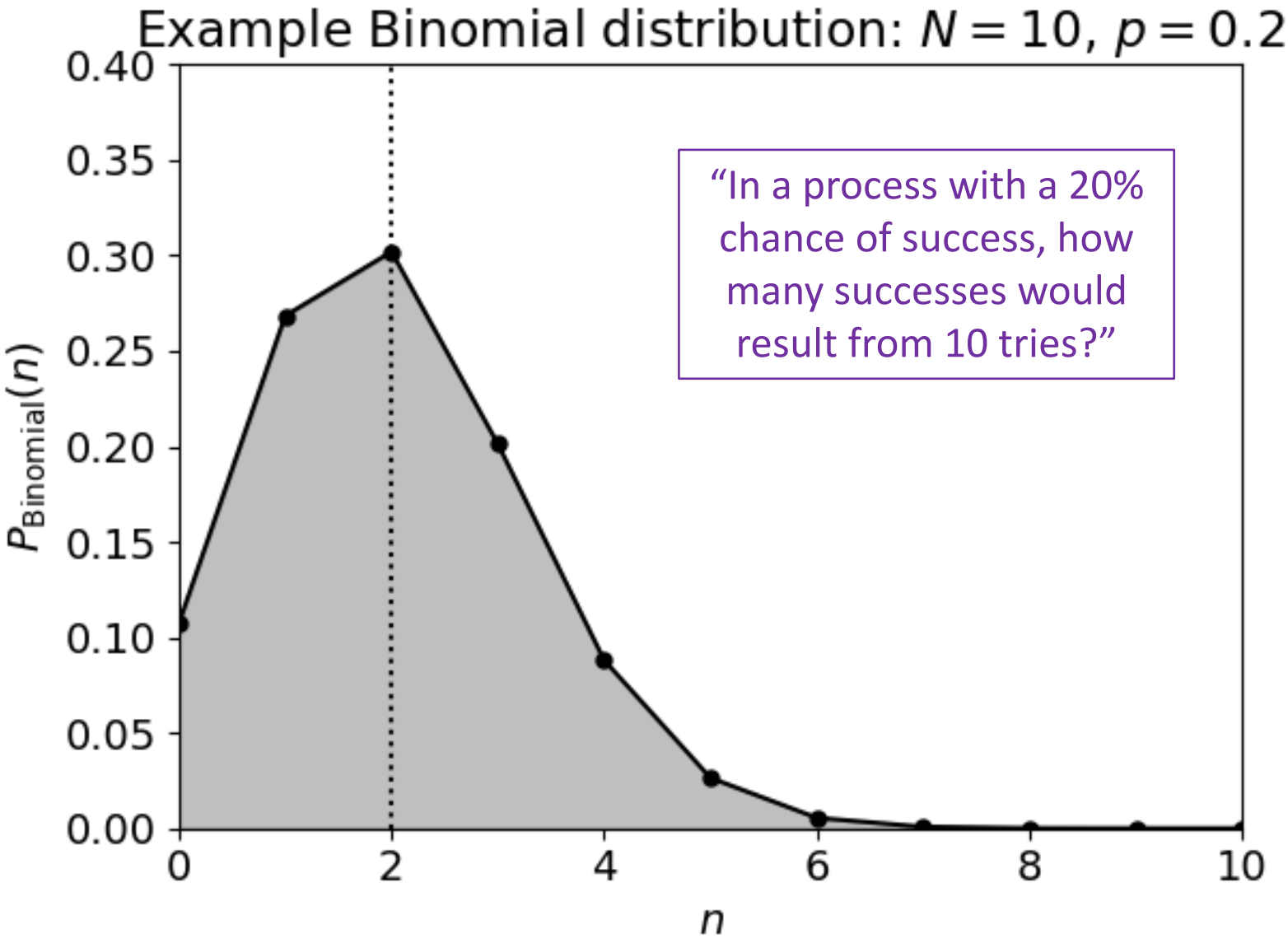
- *Example: tossing a coin*

- If we have N trials, and the probability of success in each is p, then the probability of obtaining n successes is:

$$P_{\text{Binomial}}(n) = \frac{N!}{n!\,(N-n)!}\,p^n\,(1-p)^{N-n}$$

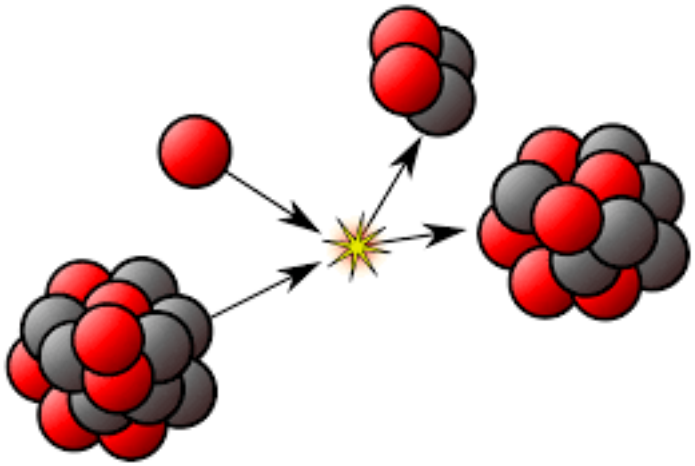
- The **mean** and **variance** of this distribution are $\bar{n} = pN$, $\text{Var}(n) = Np(1-p)$

The Binomial distribution



The Poisson distribution

- Applies to a **discrete random process where we are counting something** in a fixed interval
- *Example: radioactive decay, photons arriving at a CCD*

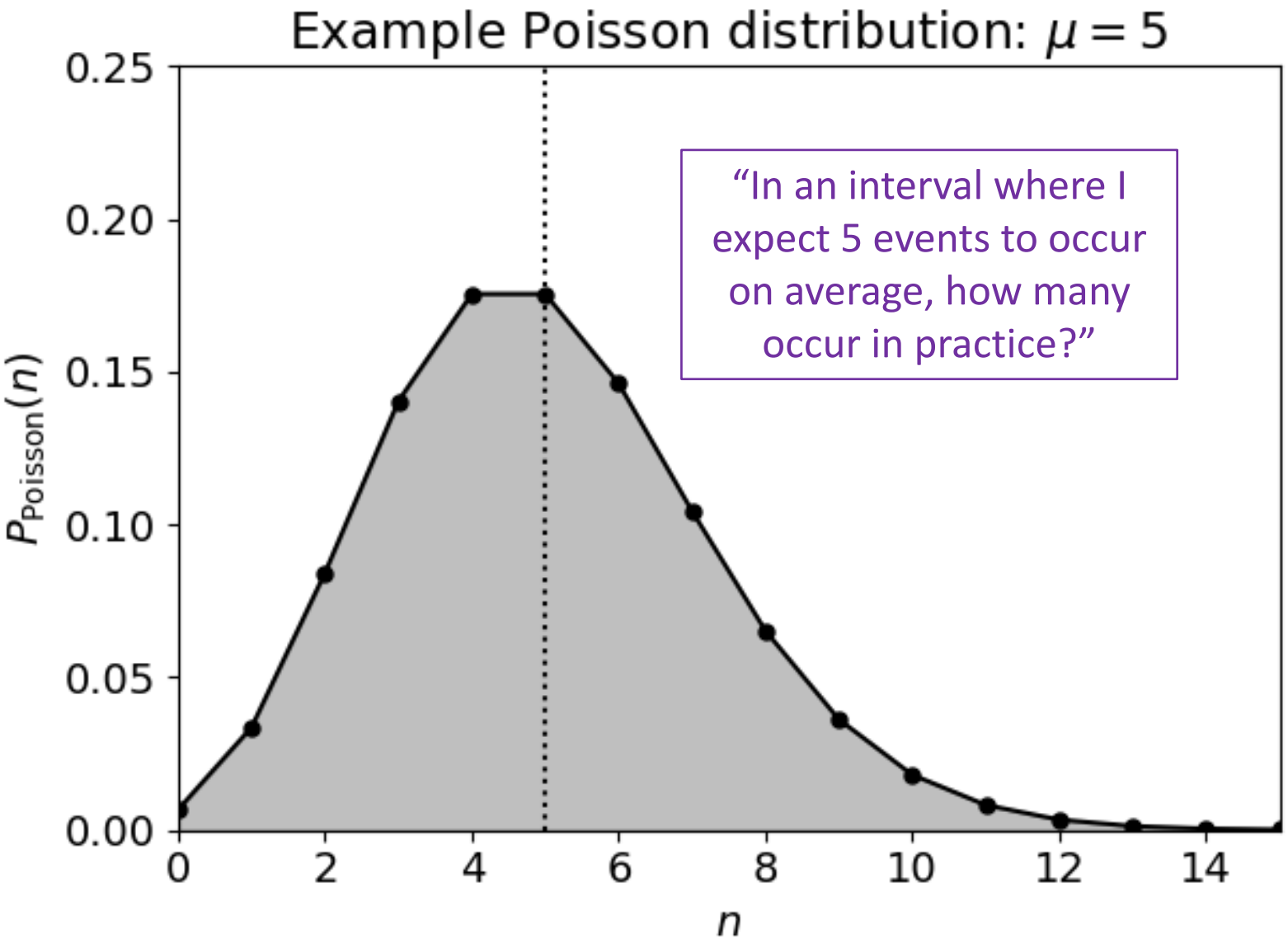


- If the mean number of events expected in some interval is μ, the probability of observing n events is

$$P_{\text{Poisson}}(n) = \frac{\mu^n \, e^{-\mu}}{n!}$$

- The **mean** and **variance** of this distribution are equal, $\bar{n} = \text{Var}(n) = \mu$

The Poisson distribution



Poisson errors

- The ultimate limit to any counting experiment
- If an individual bin of data contains N events (for example, a CCD pixel contains N photons), we can use the Poisson variance $\sigma^2 = \mu$ to place a **Poisson error** $\sqrt{N}$ in that bin

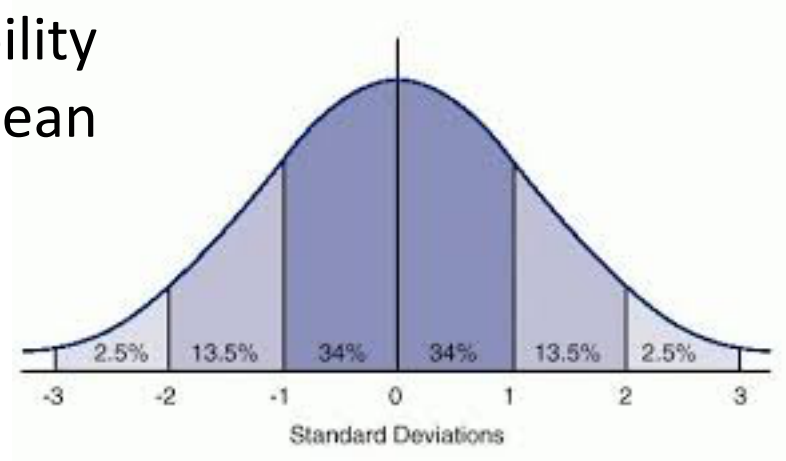
$$\text{Count} = N \pm \sqrt{N}$$

- Small print: Assumes the mean count is the observed count
- Bad approximation for low numbers (e.g. $N = 0$)
- Bad approximation if the fluctuations are dominated by other processes (e.g. read noise, galaxy clustering)

The Gaussian distribution

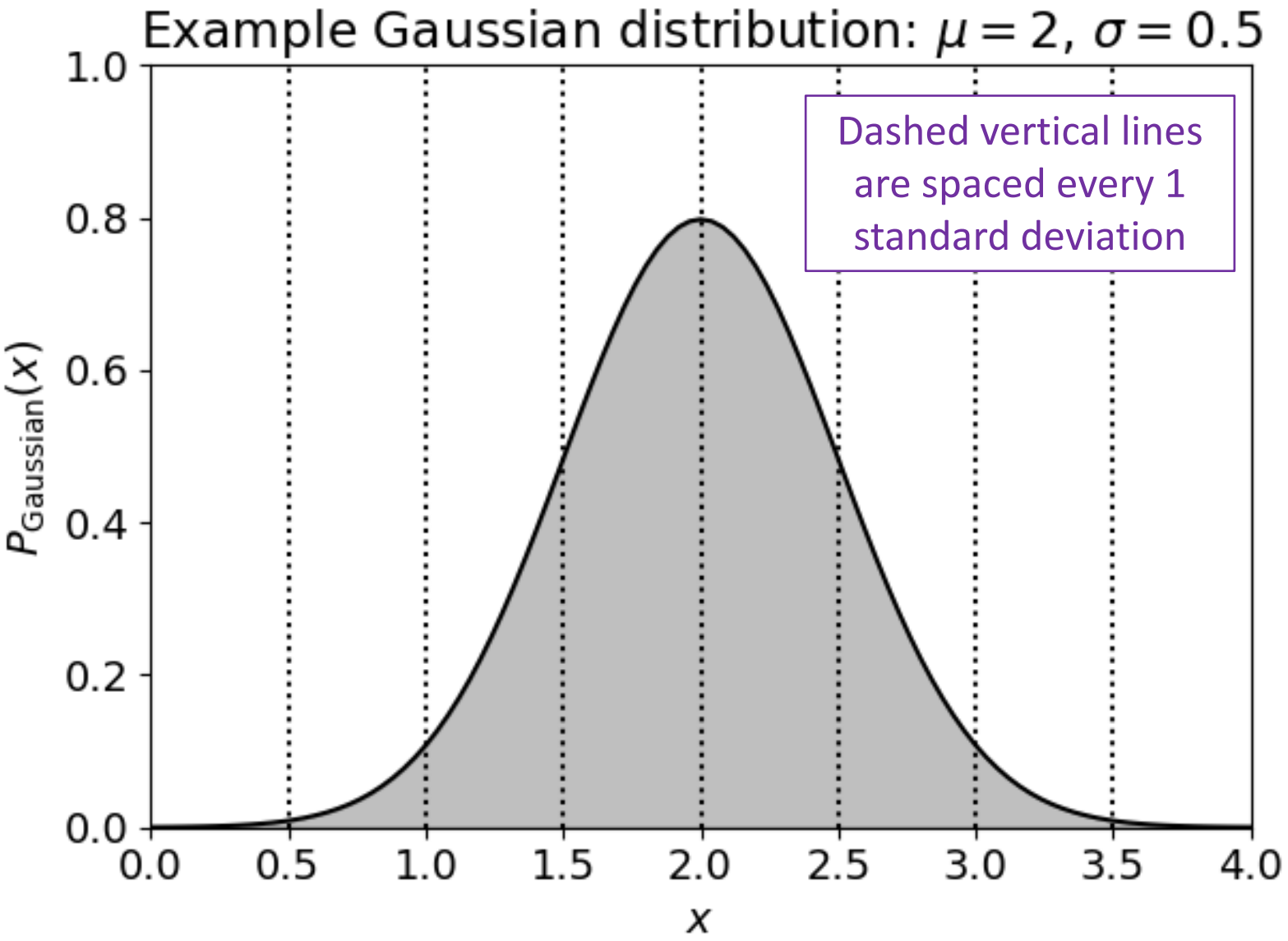
- The **Gaussian** (or "**normal**") probability distribution for a variable x, with mean μ and standard deviation σ is:

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



- *Why is this such an ubiquitous and important probability distribution?*
- It is the **high-N limit** for the Binomial and Poisson distributions
- The **central limit theorem** says that if we average together variables drawn many times from any probability distribution, the resulting average will follow a Gaussian!

The Gaussian distribution



Confidence regions and tails

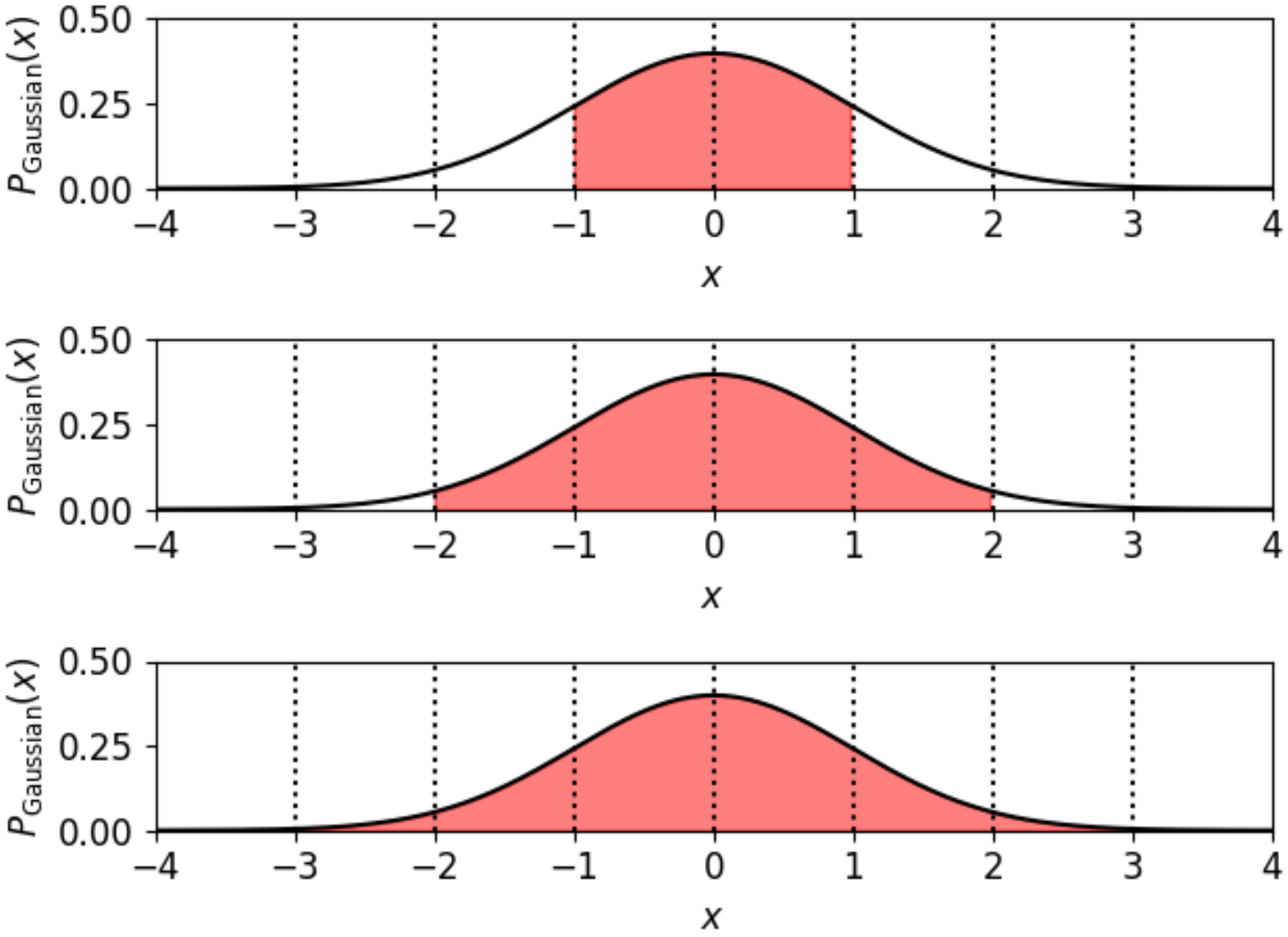
- The **Gaussian** (or "**normal**") probability distribution for a variable x, with mean μ and standard deviation σ is:

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

- The probability contained within $\pm 1,2,3$ standard deviations is $(68.27, 95.45, 99.73)\%$ (etc.)

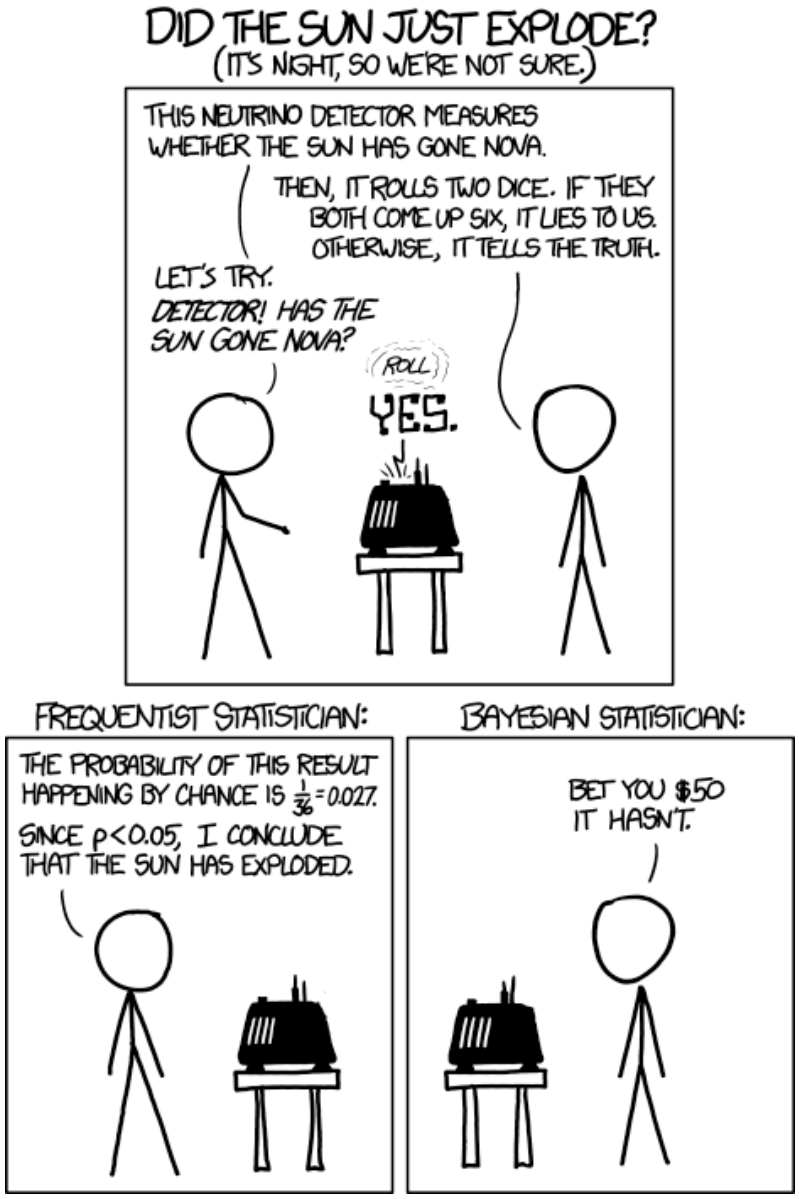
- This is often used as **shorthand for the confidence** of a statement: e.g., 3-σ confidence implies that the statement is expected to be true with a probability of 99.73%

Confidence regions and tails



Frequentist and Bayesian frameworks

- In the framework of statistics, we will often hear about **"Frequentist"** or **"Bayesian"** methods. In the next few slides we'll discuss what this means.
- Neither framework is "right" or "wrong", as such
- *As usual with statistics, it comes down to the question we want to answer …*



Credit: xkcd.com

Frequentist and Bayesian frameworks

- **Frequentist statistics** assign probabilities to a measurement, i.e. they determine $\boldsymbol{P(\mathbf{data}|\mathbf{model})}$
- We are defining probability by imaging a series of hypothetical experiments, repeatedly sampling the population (which have not actually taken place)
- *Philosophy of science: we attempt to "rule out" or falsify models, if* $P(data|model)$ *is too small*

Assuming these dice are unbiased, what is the probability of rolling different values?

Frequentist and Bayesian frameworks

- **Bayesian statistics** assign probabilities to a model, i.e. they give us tools for calculating $\boldsymbol{P(\text{model}|\text{data})}$

- We update the model probabilities in the light of each new dataset (rather than imagining many hypothetical experiments)

- *Philosophy of science: we do not "rule out" models, just determine their relative probabilities*

Assuming I roll a particular spread of different values, what is the probability of the dice being unbiased?

Frequentist and Bayesian frameworks

- The concept of **conditional probability** is central to understanding Bayesian statistics
- $P(A|B)$ means **"the probability of A on the condition that B has occurred"**
- Adding conditions makes a huge difference to evaluating probabilities
- On a randomly-chosen day in CAS, $P(\text{free pizza}) \sim 0.2$
- $P(\text{free pizza}|\text{Monday}) \sim 1$, $P(\text{free pizza}|\text{Tuesday}) \sim 0$

Frequentist and Bayesian frameworks

- The important formula for relating conditional probabilities is **Bayes' theorem**:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

(Obligatory portrait of the Reverend Bayes!)

- Small print: this formula can be derived by just writing down the joint probability of both A and B in 2 ways:

$$P(A \cap B) = P(A|B)\,P(B) = P(B|A)\,P(A)$$

- Re-writing Bayes' theorem for science:

$$P(\text{model}|\text{data}) = \frac{P(\text{data}|\text{model})\,P(\text{model})}{P(\text{data})}$$

Worked example

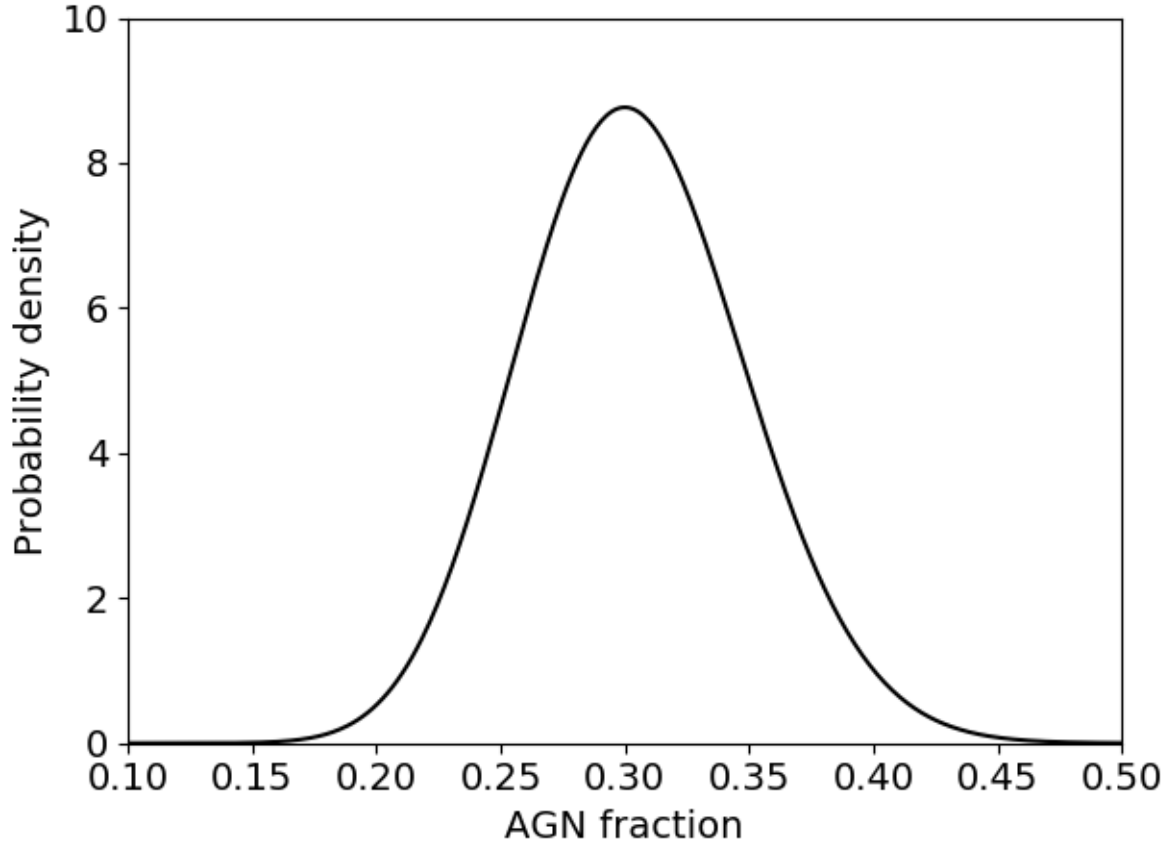
- *I observe 100 galaxies, 30 of which are AGN. What is the best estimate of the AGN fraction and its error?*
- **Solution 1**: Estimate AGN fraction $p = \frac{N_{AGN}}{N_{total}} = \frac{30}{100} = 0.3$
- There are 2 possible outcomes ("AGN" or "not an AGN") so the **binomial distribution** applies
- Estimate the error in N_{AGN} as the standard deviation in the binomial distribution $= \sqrt{N_{total}\, p(1-p)} = \sqrt{100 \times 0.3 \times 0.7} = 4.6$, so error in $p = \frac{4.6}{100} = 0.046$
- Answer: $\boldsymbol{p = 0.3 \pm 0.046}$

Worked example

- *I observe 100 galaxies, 30 of which are AGN. What is the best estimate of the AGN fraction and its error?*
- **Solution 2**: Use Bayes' theorem $P(p|D) \propto P(D|p)\, P(p)$
- $P(p|D)$ is the probability distribution of p given the data D, the quantity we aim to determine
- $P(D|p)$ is the probability of the data for a given value of p, which is given by the Binomial distribution as $P_{Binomial}(n = 30|N = 100, p)$
- $P(p)$ is the prior in p, which we take as a uniform distribution between $p = 0$ and $p = 1$
- Determining $P(p|D)$ and normalising we obtain …

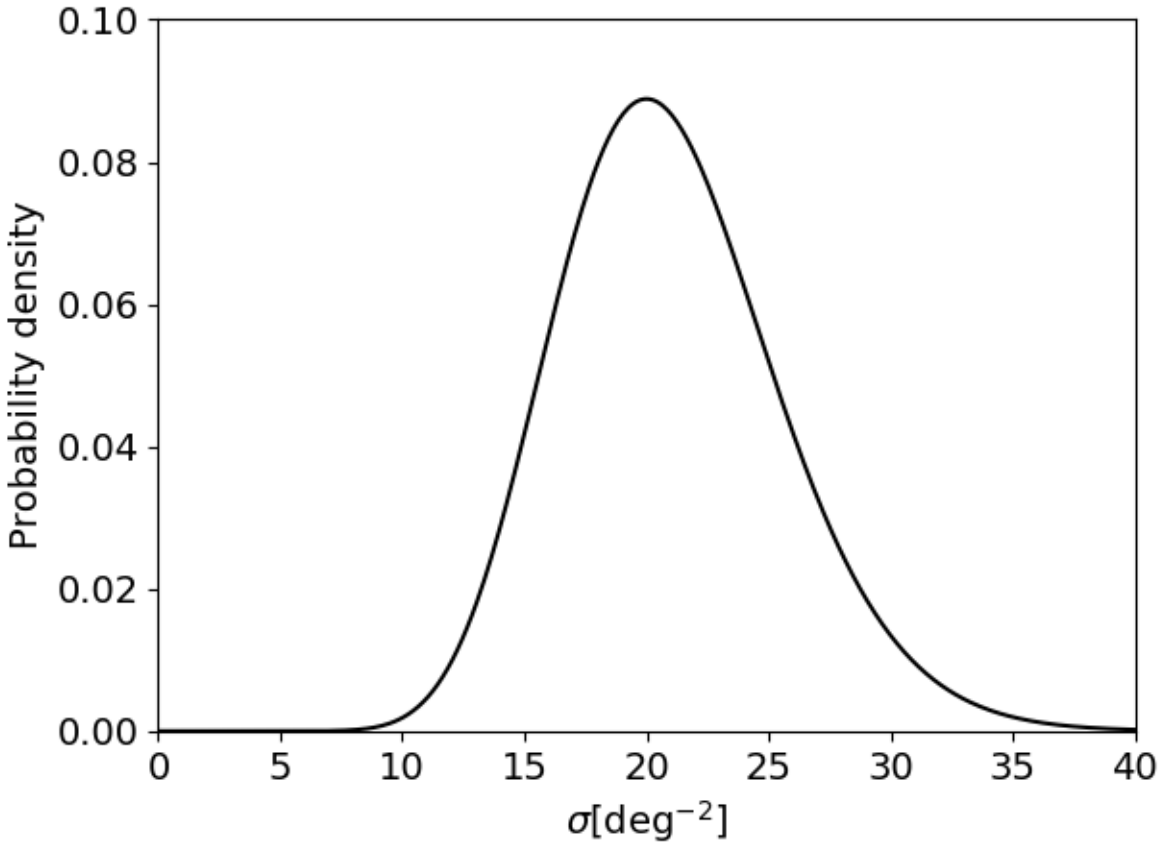
Worked example

- *I observe 100 galaxies, 30 of which are AGN. What is the best estimate of the AGN fraction and its error?*



Activity

- *A survey of area $A = 1\ \mathrm{deg}^2$ finds $N = 20$ quasars. What is the number of quasars per square degree, σ?*



Monte Carlo simulations

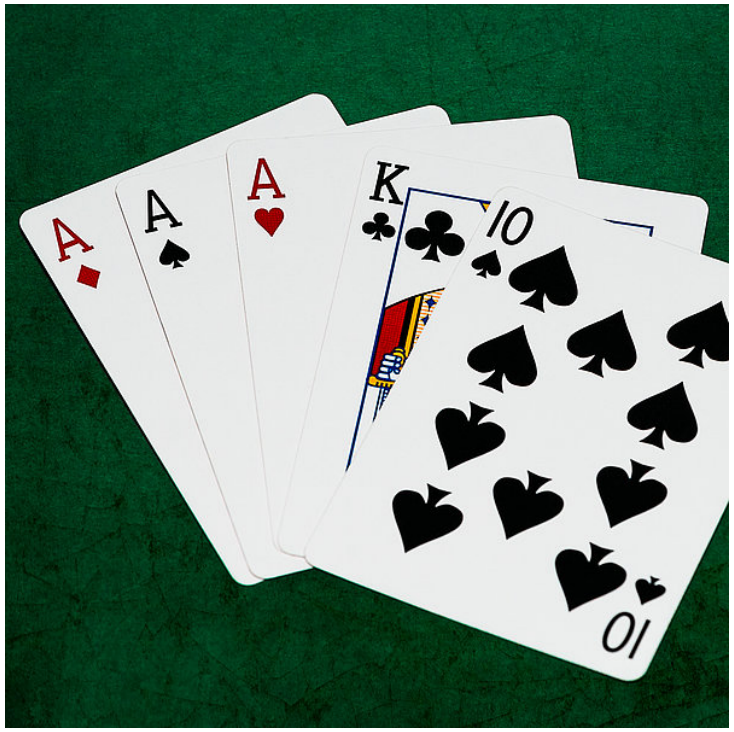
- A **Monte Carlo simulation** is a computer model of an experiment in which many random realizations of the results are created and analysed like the real data

Monte Carlo simulations

- A **Monte Carlo simulation** is a computer model of an experiment in which many random realizations of the results are created and analysed like the real data
- ***This is the most useful statistical tool you'll learn!***
- It allows us to determine the statistics of a problem **without** any analytic calculations (if we can model it)
- Statistical errors can be obtained from the **distribution of fitted parameters** over the realizations
- Systematic errors can be explored by comparing the mean fitted parameters to their **known input values**

Activity: Monte Carlo methods

- Solve the following problem by Monte Carlo methods: *I'm dealt 5 playing cards from a normal deck (i.e. 13 different values in 4 suits). What is the probability of obtaining "three of a kind" (i.e. 3 of my 5 cards having the same value?)*

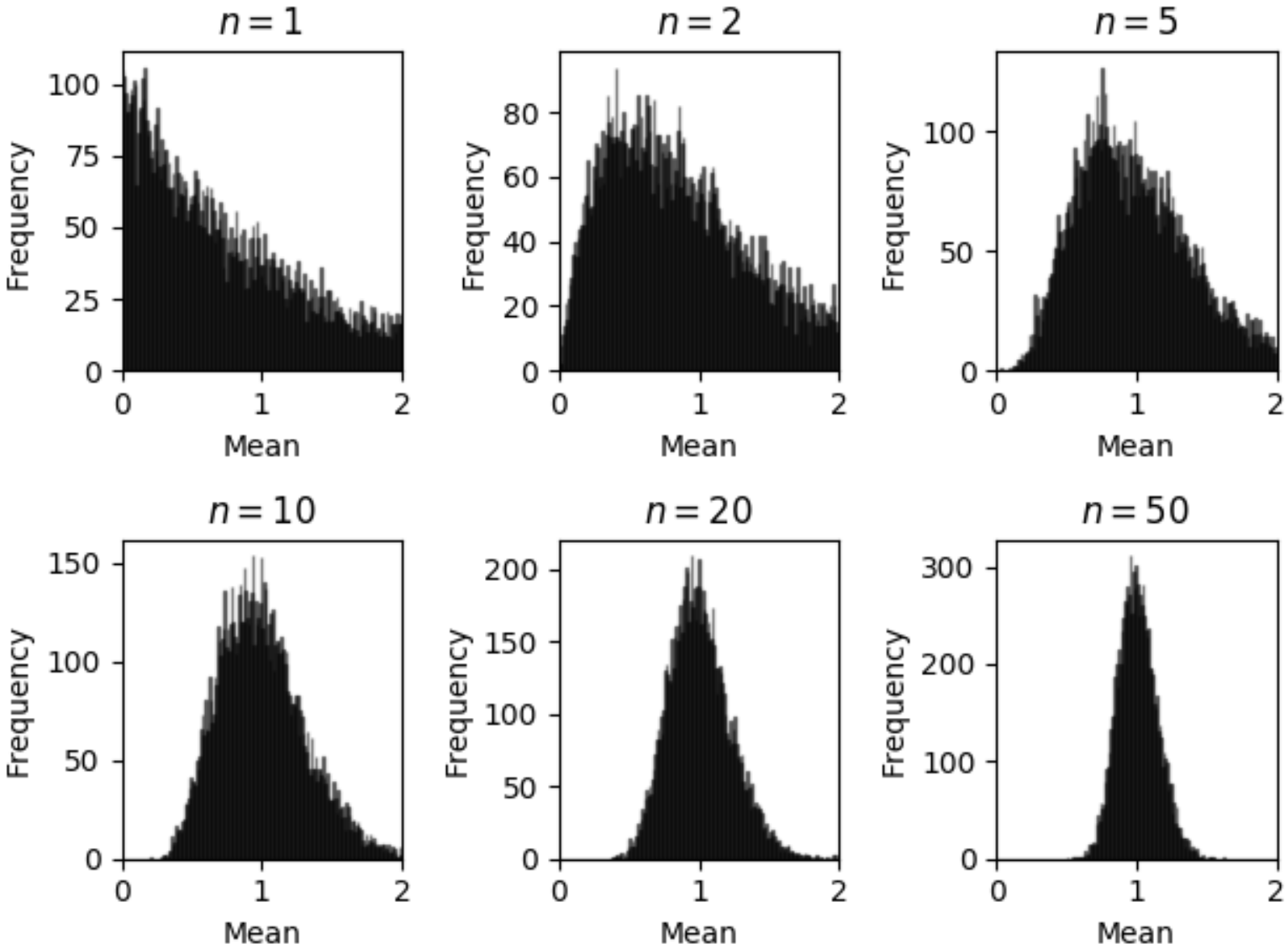


Activity: central limit theorem

- Write a code that draws n values of x from an exponential distribution $P(x) \propto e^{-x}$ (where $0 < x < \infty$), and computes their arithmetic mean μ. Repeat this process m times, and plot the probability distribution of μ across the m realisations. Run this experiment for values $n = 1, 2, 5, 10, 20, 50$.

- Hint: to do a single draw, select a uniform random number y in the range $0 < y < 1$, then $x = -\ln y$ [why does this work?]

Activity: central limit theorem



Summary

At the end of this class you should be able to …

- … determine summary statistics for datasets and their errors
- … optimally combine data
- ... apply probability distributions for Gaussian, Binomial and Poisson statistics
- … compare the Frequentist and Bayesian frameworks for statistical analysis
- … solve statistical problems using Monte Carlo techniques