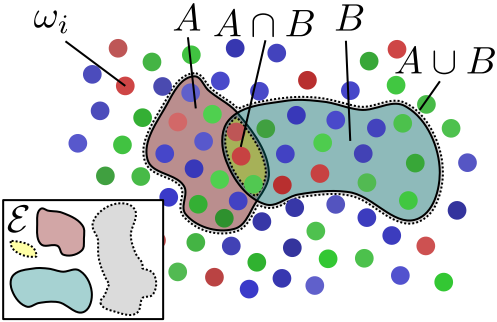
math recap!

Everything you Need to Know about Probability and Measure

MSE 421 - Ceriotti

Events and sets of events

- Consider an abstract and general notation to characterize the occurrence of an *event*, ω. ω may be e.g. "the train arrives on time", or "this molecule undergoes dissociation within one minute".
- We can then consider *sets of events* A, e.g. $A = \{\omega_1, \omega_2\}$. The set of events sets $\mathcal{E}$ must be closed under the set union and intersection operations ($A_1, A_2 \in \mathcal{E} \Rightarrow A_1 \cup A_2 \in \mathcal{E}, A_1 \cap A_2 \in \mathcal{E}$). $\mathcal{E}$ also contains an empty set $\emptyset$ and the set of all events Ω.
- In a physical setting, events may refer to the value of discrete variables (that can take on a countable number of values) or continuous variables (for which events are always to be intended as "being in a small neighborhood of a prescribed value").



Probability Axioms

- A probability is a function $P:\mathcal{E}\to\mathbb{R}^+$ that satisfies the following axioms:
  1. $\forall A\in\mathcal{E}, P(A)\geq 0$
  2. $P(\Omega)=1$
  3. For any *countable* collection of sets $\{A_i\}$ that are non-overlapping (such that $A_i\cap A_j=\emptyset$)

$$P(A_1\cup A_2\ldots)=\sum_i P(A_i)$$

- From these axioms it follows that
  1. If $\bar{A}$ is the complement of A ($A\cup\bar{A}=\Omega$ and $A\cap\bar{A}=\emptyset$) [consider axioms 2 and 3]

$$P\left(\bar{A}\right)=1-P(A)$$

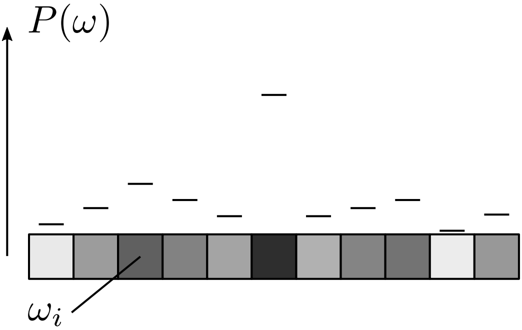
  2. The empty set has probability zero, $P(\emptyset)=0$ [consider axiom 3 and see that $P(A)=P(A\cup\emptyset)=P(A)+P(\emptyset)$]

What does this mean?

- "Intuitive" understanding of probability tends to be ill-defined: if we observe an event N *independent* times, the outcome ω will belong to the set A a number $NP(A)$ times, provided N is "*large enough*"
- Individual events are mutually exclusive (cannot happen together).
  - For *discrete* events (toss of a coin, sum of two dice, . . .) one can assign probabilities to individual events.
  - For phenomena that can take a *continuous* value, we can only define the probability of ω falling within a range of values

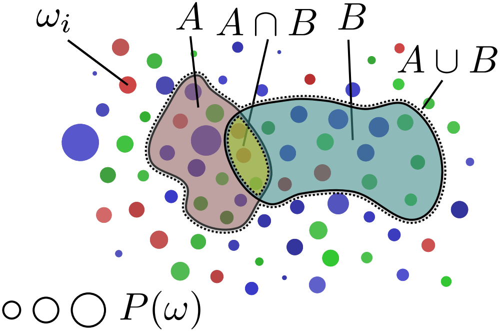
What does this mean?

- "Intuitive" understanding of probability tends to be ill-defined: if we observe an event N *independent* times, the outcome ω will belong to the set A a number $NP(A)$ times, provided N is "*large enough*"
- Individual events are mutually exclusive (cannot happen together).
  - For *discrete* events (toss of a coin, sum of two dice, . . .) one can assign probabilities to individual events.
  - For phenomena that can take a *continuous* value, we can only define the probability of ω falling within a range of values



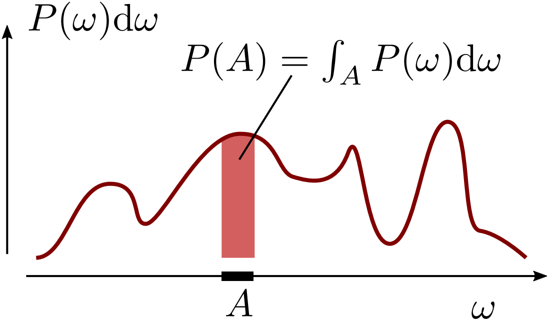
What does this mean?

- “Intuitive” understanding of probability tends to be ill-defined: if we observe an event *N independent* times, the outcome ω will belong to the set *A* a number $NP(A)$ times, provided *N* is “*large enough*”
- Individual events are mutually exclusive (cannot happen together).
  - For *discrete* events (toss of a coin, sum of two dice, . . .) one can assign probabilities to individual events.
  - For phenomena that can take a *continuous* value, we can only define the probability of ω falling within a range of values



What does this mean?

- "Intuitive" understanding of probability tends to be ill-defined: if we observe an event *N independent* times, the outcome ω will belong to the set *A* a number $NP(A)$ times, provided *N* is "*large enough*"
- Individual events are mutually exclusive (cannot happen together).
  - For *discrete* events (toss of a coin, sum of two dice, . . .) one can assign probabilities to individual events.
  - For phenomena that can take a *continuous* value, we can only define the probability of ω falling within a range of values



$P(\omega)\mathrm{d}\omega$
$P(A)=\int_A P(\omega)\mathrm{d}\omega$
A
ω

Joint, marginal and conditional probabilities

- Consider two event sets $A \in \mathcal{E}$, $B \in \mathcal{E}'$. We can define the **joint** probability $P(A, B)$ as the probability that two events compatible with the sets A and B both occur. Formally, one can think this in terms of more complex events living in the product space $\mathcal{E} \times \mathcal{E}'$, but we can ignore the subtlety.

$$P(A, B) = P(\omega \in A \text{ and } \omega' \in B)$$

- Given a joint probability we can recover the probability of individual events by considering *marginal* probability, e.g.

$$P(A, \Omega') = P(A), \qquad P(\Omega, B) = P(B)$$

- We can define *conditional* probabilities as the probability of an event set A conditional on knowledge that event set B occurs

$$P(A|B) = P(\omega \in A \text{ knowing that } \omega' \in B)$$

Marginalization of a joint probability

- Say you could decompose Ω into a (countable) set of disjoint components $A_1, \ldots$, i.e. $\cup_i A_i = \Omega$, $A_i \cap A_j = \emptyset$. We can use axiom 3 to obtain marginals by summing over all possible events

$$P\left(B\right) = P\left(\Omega, B\right) = \sum_i P\left(A_i, B\right)$$

- Generalizing, we can reduce a complex joint probability by marginalization:

$$P\left(B, C, D\right) = \sum_i P\left(A_i, B, C, D\right)$$

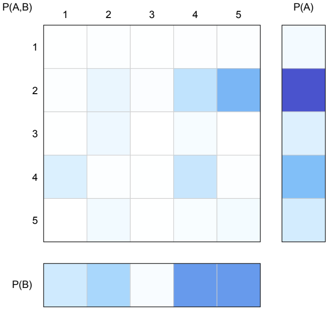
Marginalization of a joint probability

- Say you could decompose Ω into a (countable) set of disjoint components $A_1, \ldots$, i.e. $\cup_i A_i = \Omega$, $A_i \cap A_j = \emptyset$. We can use axiom 3 to obtain marginals by summing over all possible events

$$P(\omega_2) = \sum_i P(\omega_i, \omega_2) \qquad P(y) = \int P(x, y)\, dx$$

- Generalizing, we can reduce a complex joint probability by marginalization:

$$P(\omega_2, \omega_3) = \sum_i P(\omega_i, \omega_2, \omega_3) \qquad P(y, z) = \int P(x, y, z)\, dx$$



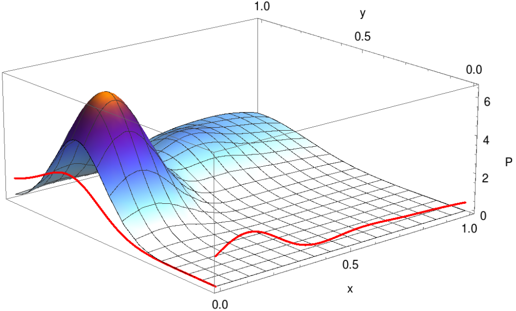
Marginalization of a joint probability

- Say you could decompose Ω into a (countable) set of disjoint components $A_1, \ldots$, i.e. $\cup_i A_i = \Omega$, $A_i \cap A_j = \emptyset$. We can use axiom 3 to obtain marginals by summing over all possible events

$$P(\omega_2) = \sum_i P(\omega_i, \omega_2) \qquad P(y) = \int P(x, y)\, \mathrm{d}x$$

- Generalizing, we can reduce a complex joint probability by marginalization:

$$P(\omega_2, \omega_3) = \sum_i P(\omega_i, \omega_2, \omega_3) \qquad P(y, z) = \int P(x, y, z)\, \mathrm{d}x$$



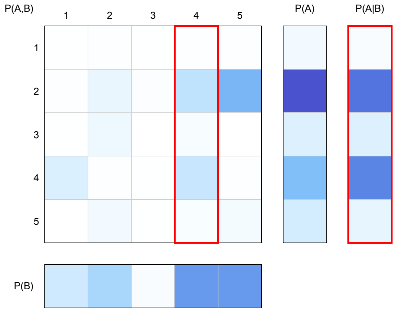
Conditional probability and Bayes' formula

- The conditional probability can be related to the joint probability

$$\boxed{P(A|B) = P(A,B)/P(B)} \qquad \text{Bayes' formula}$$

  - Basically this amounts at renormalizing the joint probability based on the knowledge we have of the outcome of one of the events
  - If the events are *independent*, knowledge of the outcome B does not give us information, so $P(A|B) = P(A)$. We can then "define" independent events as those for which

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



Bayes' formula in action

Why is it hard to develop tests for rare diseases?

1. Linus developed a test that gives no false negatives but has a $1/1'000$ probability of false positives [If you have the disease, the test will certainly catch it. If you don't have the disease, there is one probability in 1000 that the test is wrong and says you have it]
2. The disease has an incidence in the general population of $1/1'000'000$ [if you pick a random human being, only one in a million actually has the disease]
3. Bill takes the test. Ouch, he's positive! Shall Bill freak out about his test?

Bayes' formula in action

Why is it hard to develop tests for rare diseases?

1. Linus developed a test that gives no false negatives but has a $1/1'000$ probability of false positives [If you have the disease, the test will certainly catch it. If you don't have the disease, there is one probability in 1000 that the test is wrong and says you have it]
2. The disease has an incidence in the general population of $1/1'000'000$ [if you pick a random human being, only one in a million actually has the disease]
3. Bill takes the test. Ouch, he's positive! Shall Bill freak out about his test?

- A: the subject is affected by the disease; B: the test is positive
- $P\left(A|B\right)$ probability that a person that comes out positive has the disease
  - $P\left(A\right) = 1/1'000'000$, $P\left(B|A\right) = 1$
  - $P\left(\bar{B}|\bar{A}\right) = 999/1'000$, $P\left(B|\bar{A}\right) = 1/1'000$,
  - $P\left(B\right) = P\left(A, B\right) + P\left(\bar{A}, B\right) = P\left(B|A\right) P\left(A\right) + P\left(B|\bar{A}\right) P\left(\bar{A}\right) =$ $10^{-6} + 10^{-3}\left(1 - 10^{-6}\right) \approx 10^{-3}$
  - $P\left(A|B\right) = P\left(A, B\right) / P\left(B\right) = P\left(B|A\right) P\left(A\right) / P\left(B\right) = 10^{-6}/10^{-3} = 10^{-3}$

Random variables and their functions

- An event can often be associated with the value of one or more quantities, $X_i(\omega)$.
- When these values completely determine the event, we can as well label it with the value of these, and write $P(\omega) = P(\mathbf{X})$.
- For *continuous* variables, we have to define probability densities, that only make sense when written in an integral form, e.g. $P(X)\,\mathrm{d}X$
- We can also of course compute functions of random variables, e.g. $f(\mathbf{X})$
  - Function of random variables are in turns random variables, characterized by their own distribution [if $f(X)$ is monotonic, $P(f)\,\mathrm{d}f = P\left(X^{-1}(f)\right) f'(X)\,\mathrm{d}X$]

Averages, moments, cumulants

- We can characterize a random variable by a series of averages performed over the probability distribution

$$\langle X^n \rangle = \sum_X X^n P(X) \, , \int X^n P(X) \, \mathrm{d}X$$

  - Moments of high order contain information on lower moments, because $\left\langle X^{2n} \right\rangle \geq \langle X^n \rangle^2$
  - We can define combinations of moments that (to an extent) eliminate this dependence, called *cumulants*. These can be complex, but the first is equal to the average, and the second to the *variance*

$$\mathrm{var}(X) = \left\langle (X - \langle X \rangle)^2 \right\rangle = \left\langle X^2 \right\rangle - \langle X \rangle^2$$

- When there are multiple variables, it is useful to define a *covariance*, that gives information on how correlated are the two variables

$$\left\langle X_i, X_j \right\rangle = \left\langle (X_i - \langle X_i \rangle) \left(X_j - \left\langle X_j \right\rangle \right) \right\rangle = \left\langle X_i X_j \right\rangle - \langle X_i \rangle \left\langle X_j \right\rangle$$

  - If the variables are independent in pairs, $\langle X_i, X_j \rangle = \mathrm{var}(X_i) \, \delta_{ij}$ [the opposite is not true, two variables can be correlated and have zero covariance, think $\cos\theta$ and $\sin\theta$]

Means, correlations & c

- Let's say we have N evaluations (independent or not) of a random variable $X\left(i\right)$ [we label for simplicity successive realizations of the event ω with an integer index i]
- The mean of the set of events is itself a random variable, $\bar{X}_N = \frac{1}{N}\sum_i X\left(i\right)$
  - Its average equals the average of X:

$$\left\langle \bar{X}_N \right\rangle = \frac{1}{N}\sum_i \left\langle X\left(i\right)\right\rangle = \left\langle X\right\rangle$$

  - Its variance depends on the correlation between different occurrences

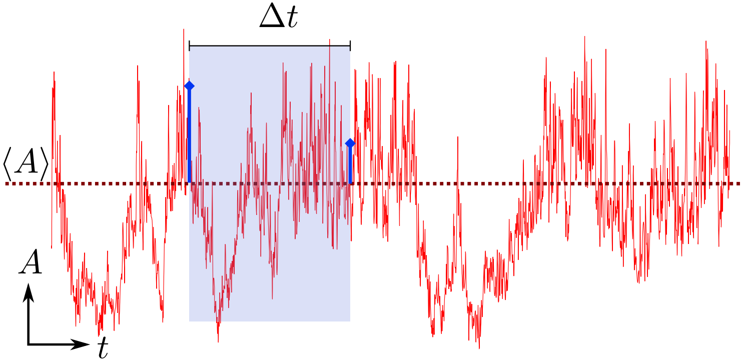
$$\operatorname{var}\left(\bar{X}_N\right) = \left\langle \bar{X}_N, \bar{X}_N \right\rangle = \frac{1}{N^2}\sum_{ij} \left\langle X\left(i\right), X\left(j\right)\right\rangle$$

  - *If we assume uncorrelated samples* $\left\langle X\left(i\right), X\left(j\right)\right\rangle = \operatorname{var}\left(X\right)\delta_{ij}$

$$\operatorname{var}\left(\bar{X}_N\right) = \frac{\operatorname{var}\left(X\right)}{N^2}\sum_{ij}\delta_{ij} = \frac{\operatorname{var}\left(X\right)}{N}$$

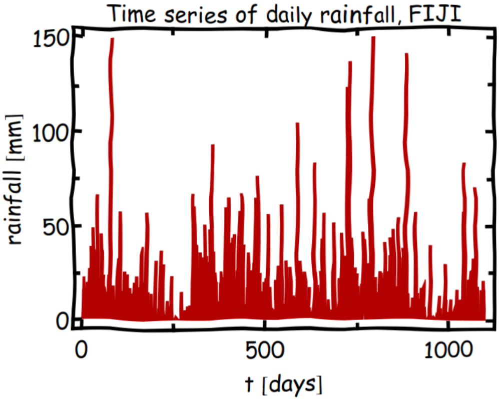
Autocorrelation function

- Often samples come from a time series, and different samples are correlated so $c_{XX}(j) = \langle X(i), X(i+j)\rangle / \mathrm{var}(X)$ will be a decaying *autocorrelation function* of j
  - With correlated samples, the error in the mean will decay more slowly ~ $\mathrm{var}(X)(N/\nu)^{-1}$, where $\nu = \sum_j c_{XX}(j)$ is the *autocorrelation time*
- Physically $c_{XX}(j)$ says how fast fluctuations from the mean are forgotten



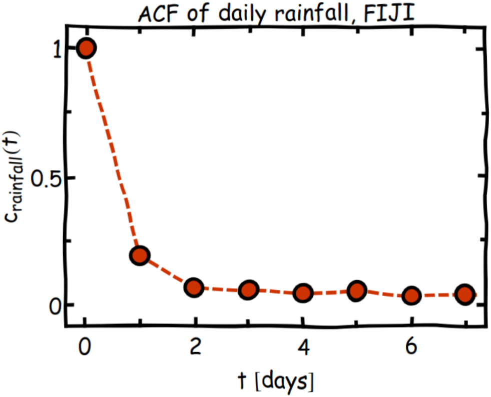
Autocorrelation function

- Often samples come from a time series, and different samples are correlated so $c_{XX}(j) = \langle X(i), X(i+j)\rangle / \mathrm{var}(X)$ will be a decaying *autocorrelation function* of j
    - With correlated samples, the error in the mean will decay more slowly ~ $\mathrm{var}(X)(N/\nu)^{-1}$, where $\nu = \sum_j c_{XX}(j)$ is the *autocorrelation time*
- Physically $c_{XX}(j)$ says how fast fluctuations from the mean are forgotten



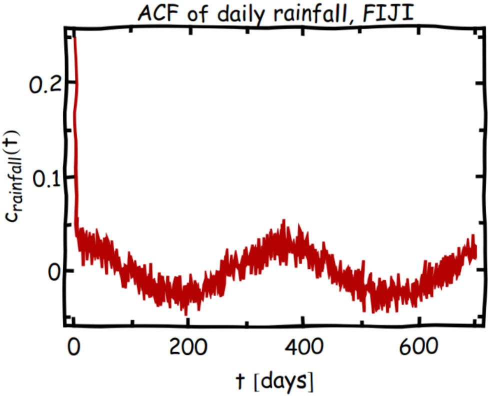
Autocorrelation function

- Often samples come from a time series, and different samples are correlated so $c_{XX}(j) = \langle X(i), X(i+j) \rangle / \mathrm{var}(X)$ will be a decaying *autocorrelation function* of j
  - With correlated samples, the error in the mean will decay more slowly ~ $\mathrm{var}(X)(N/\nu)^{-1}$, where $\nu = \sum_j c_{XX}(j)$ is the *autocorrelation time*
- Physically $c_{XX}(j)$ says how fast fluctuations from the mean are forgotten



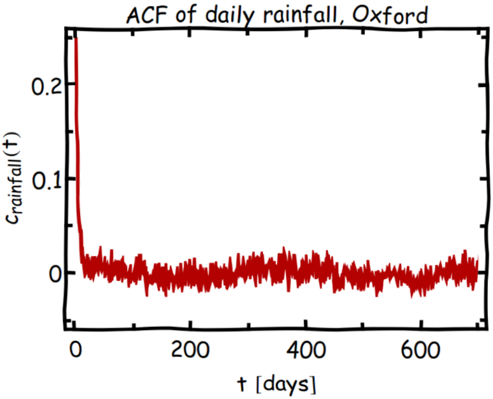
Autocorrelation function

- Often samples come from a time series, and different samples are correlated so $c_{XX}(j) = \langle X(i), X(i+j)\rangle / \mathrm{var}(X)$ will be a decaying *autocorrelation function* of j
  - With correlated samples, the error in the mean will decay more slowly ~ $\mathrm{var}(X)(N/\nu)^{-1}$, where $\nu = \sum_j c_{XX}(j)$ is the *autocorrelation time*
- Physically $c_{XX}(j)$ says how fast fluctuations from the mean are forgotten



Autocorrelation function

- Often samples come from a time series, and different samples are correlated so $c_{XX}(j) = \langle X(i), X(i+j)\rangle / \mathrm{var}(X)$ will be a decaying *autocorrelation function* of j
  - With correlated samples, the error in the mean will decay more slowly ~ $\mathrm{var}(X)(N/\nu)^{-1}$, where $\nu = \sum_j c_{XX}(j)$ is the *autocorrelation time*
- Physically $c_{XX}(j)$ says how fast fluctuations from the mean are forgotten



Characteristic function

- Given a probability function $P(X)$ we can define the characteristic function - which is just its Fourier transform

$$\phi(s) = \langle \exp \mathrm{i}sX \rangle = \int P(X)\, e^{\mathrm{i}sX} \mathrm{d}X$$

  - If P is normalized, $\phi(0) = 1$
  - From the properties of the FT, $\phi^{(n)}(0) = \mathrm{i}^n \langle X^n \rangle$ [this is why $\phi(s)$ is also know as the *moment generating function*]

- Consider the characteristic function associated with the joint sum of two independent variables

$$\phi_{1+2}(s) = \int P(X_1, X_2)\, e^{\mathrm{i}s(X_1+X_2)} \mathrm{d}X_1 \mathrm{d}X_2 \underset{\text{if independent}}{=} \phi_1(s)\, \phi_2(s)$$

- This is very useful to predict the distribution of the combination of independent variables, such as the mean!

$$\phi_{\bar{X}_N}(s) = \prod_i \phi_{X(i)}(s/N) = \left[\phi_X(s/N)\right]^N$$

Central limit theorem

- Consider a distribution with finite moments, assuming for simplicity zero mean and unit variance. Taylor expand it

$$\phi_X(s) \approx 1 - \frac{1}{2}s^2 + \mathcal{O}\left(s^3\right)$$

- Now consider the characteristic function of the mean

$$\phi_{\bar{X}_N}(s) = \left[\phi_X(s/N)\right]^N \approx \left[1 - \frac{1}{N}\frac{1}{2}\frac{s}{N}^2 + \mathcal{O}\left(\left(\frac{s}{N}\right)^3\right)\right]^N \xrightarrow[N\to\infty]{} e^{-\frac{s^2}{2N}}$$

  - Inverting the definition of the characteristic function, one gets $P\left(\bar{X}_N\right) \propto e^{-N\bar{X}_N^2/2}$, i.e. the mean is Gaussian distributed with a variance $1/N$

- Note that the value of the variance depends on correlations between variables, but the Gaussian nature is guaranteed provided that the moments of $P(X)$ do not grow too quickly. *Regardless of the details of the distribution of individual samples, the mean of a large number of independent terms has a Gaussian distribution.*

Recall that $\lim_{N\to\infty}\left(1+\frac{x}{N}\right)^N = e^x$.

Binomial distribution

- Given a binary event A which has a probability $P(t) = q$ of being true and $P(f) = 1 - q$ of being false, what is the probability that given n independent instances of the event, m will be true?
  - Since the events are assumed independent,

$$P(A(1), A(2), \ldots A(n)) = \prod_i P(A(i))$$

- We do not specify the order of realizations, so e.g. *ttft* and *fttt* are equally valid realizations of three positive outcomes. So summing over all possible realizations

$$B(n,m) = q^m (1-q)^{n-m} \frac{n}{m} \cdot \frac{n-1}{m-1} \cdots \frac{n-m+1}{1} = q^m (1-q)^{n-m} \frac{n!}{m!(n-m)!}$$

  - Mean value: $\langle m \rangle = \sum_i iB(n,i) = nq, \mathrm{var}(m) = nq(1-q)$
- Verify central limit theorem, taking large n, m limit

Binomial distribution

- Given a binary event A which has a probability $P(t) = q$ of being true and $P(f) = 1 - q$ of being false, what is the probability that given n independent instances of the event, m will be true?
    - Since the events are assumed independent,

$$P(A(1), A(2), \dots A(n)) = \prod_i P(A(i))$$

- We do not specify the order of realizations, so e.g. *ttft* and *fttt* are equally valid realizations of three positive outcomes. So summing over all possible realizations

$$B(n,m) = q^m (1-q)^{n-m} \frac{n}{m} \cdot \frac{n-1}{m-1} \cdots \frac{n-m+1}{1} = q^m (1-q)^{n-m} \frac{n!}{m!(n-m)!}$$

- Mean value: $\langle m \rangle = \sum_i iB(n,i) = nq$, $\mathrm{var}(m) = nq(1-q)$

- Verify central limit theorem, taking large n, m limit

$$\ln B = m \ln q + (n-m)\ln(1-q) + n \ln n - (n-m)\ln(n-m) - m \ln m$$

$$\frac{\partial \ln B}{\partial m} = \ln \frac{q}{1-q} + \ln \frac{n-m}{m} = 0 \to m = nq$$

$$\left. \frac{\partial^2 \ln B}{\partial m^2} \right\|_{m=nq} = \frac{1}{n(1-q)} - \frac{1}{nq} = -\frac{1}{nq(1-q)}$$

Binomial distribution

- Given a binary event A which has a probability $P(t) = q$ of being true and $P(f) = 1 - q$ of being false, what is the probability that given n independent instances of the event, m will be true?
    - Since the events are assumed independent,

$$P(A(1), A(2), \ldots A(n)) = \prod_i P(A(i))$$

- We do not specify the order of realizations, so e.g. *ttft* and *fttt* are equally valid realizations of three positive outcomes. So summing over all possible realizations

$$B(n, m) = q^m (1 - q)^{n-m} \frac{n}{m} \cdot \frac{n-1}{m-1} \cdots \frac{n-m+1}{1} = q^m (1 - q)^{n-m} \frac{n!}{m!(n-m)!}$$

- Mean value: $\langle m \rangle = \sum_i iB(n, i) = nq$, $\mathrm{var}(m) = nq(1 - q)$

- Verify central limit theorem, taking large n, m limit

There There

$$\ln B \approx -\frac{(m - nq)^2}{2nq(1-q)} \leftarrow \text{Gaussian with mean } nq \text{ and variance } nq(1-q)$$

Correlations (and causation?)

- Consider two events A, B, each with multiple outcomes, $A1$, $A2$, . . .
- How to check for significance of the correlations? Null hypothesis: the two events are independent.
- χ^2 test measures extent of correlations, and their significance (p-value)
  - Measure deviation between observed joint frequencies, and those expected assuming uncorrelated outcomes
  - Compute total deviation and compare with χ^2 statistics

	swiss	italian
punctual	100	5
late	3	12

	swiss	italian
punctual	10	1
late	3	2

Correlations (and causation?)

- Consider two events A, B, each with multiple outcomes, $A1$, $A2$, . . .
- How to check for significance of the correlations? Null hypothesis: the two events are independent.
- χ^2 test measures extent of correlations, and their significance (p-value)
  - Measure deviation between observed joint frequencies, and those expected assuming uncorrelated outcomes
  - Compute total deviation and compare with χ^2 statistics

	swiss	italian	
punctual	100	5	105
late	3	12	15
	103	17	

Compute marginals

Correlations (and causation?)

- Consider two events *A*, *B*, each with multiple outcomes, *A*1, *A*2, . . .
- How to check for significance of the correlations? Null hypothesis: the two events are independent.
- χ^2 test measures extent of correlations, and their significance (p-value)
  - Measure deviation between observed joint frequencies, and those expected assuming uncorrelated outcomes
  - Compute total deviation and compare with χ^2 statistics

	swiss	italian	
punctual	100 (90.1)	5 (14.9)	105
late	3 (12.9)	12 (2.1)	15
	103	17	120

Compute expected outputs based on the marginals

Correlations (and causation?)

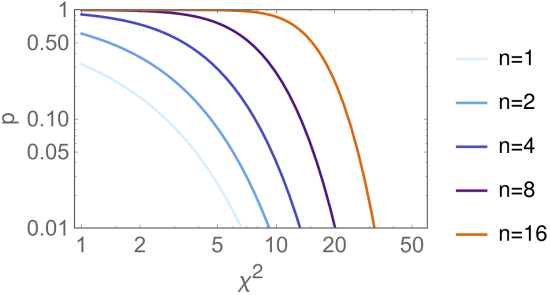
- Consider two events A, B, each with multiple outcomes, $A1$, $A2$, . . .
- How to check for significance of the correlations? Null hypothesis: the two events are independent.
- χ^2 test measures extent of correlations, and their significance (p-value)
  - Measure deviation between observed joint frequencies, and those expected assuming uncorrelated outcomes
  - Compute total deviation and compare with χ^2 statistics

$\left(E_i - O_i\right)^2 / E_i$	swiss	italian	
punctual	1.1	8.0	
late	9.2	46.7	
n=1			$\chi^2 = 65$

Compute χ^2, $\sum_i \left(E_i - O_i\right)^2 / E_i$

Correlations (and causation?)

- Consider two events A, B, each with multiple outcomes, $A1$, $A2$, . . .
- How to check for significance of the correlations? Null hypothesis: the two events are independent.
- χ^2 test measures extent of correlations, and their significance (p-value)
  - Measure deviation between observed joint frequencies, and those expected assuming uncorrelated outcomes
  - Compute total deviation and compare with χ^2 statistics



Compare with the CDF for χ^2 statistics, in this case $p = 10^{-15}$