BASIC CONCEPTS IN GENETICS AND HOME-EXERCISE 1

● Extension of the content in the course *Genetic analysis and molecular evolution* (GAME) which is the background of most students in the current course *Statistical genetics* (STAGE)

● This set of slides includes some relevant GAME-slides and home-exercises with solutions. Read them!

● New parts include home-exercises which you should solve and submit 28. November at the latest. You can submit them as handwritten to Siru, or to course Moodle. You´ll get correct answers 3. December and thus become prepared to the last course module (GWAS) during which you need understanding of these concepts.

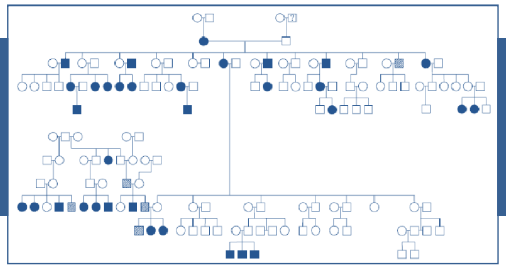
● You can work alone or as a group of two students and submit joint solutions.

● If your don´t want to work for home-exercises, that is ok. However, then you will not be very well prepared to the exam. Excellent performance with home-exercises can give you a better final grade than the one you get from the exam.

● You should not expect exam-questions from the GAME-part.

FROM GAME

INHERITANCE OF TRAITS PEDIGREE ANALYSIS



● The aim of this topic is to familiarize with Mendelian inheritance + pedigree analysis + probabilistic inference.

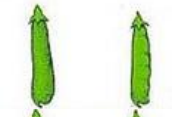
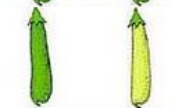
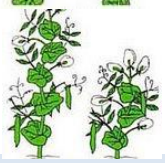
● These principles are the elementary ones behind the current “Big data” questions and data analysis.

● Inheritance of biological traits has been recognized for thousands of years: *traits go in families.*

● The first insights how inheritance of traits takes place occurred ~150 years ago when **Gregor Mendel** published the results of a series of experiments that would lay the foundation for the formal discipline of *genetics*.

● Mendel used a model experimental approach to study patterns of inheritance and derived important postulates: (1) *Genetic characters are controlled by unit factors existing in pairs in individual organisms.* (2) *When two unlike unit factors responsible for a single character are present in a single individual, one unit factor is dominant to the other, which is said to be recessive.* (3) *During the formation of gametes, the paired unit factors separate (segregate) randomly so that each gamete receives one or the other with equal probability.* (4) *During the gamete formation, segregating pairs of unit factors assort independently of each other.*

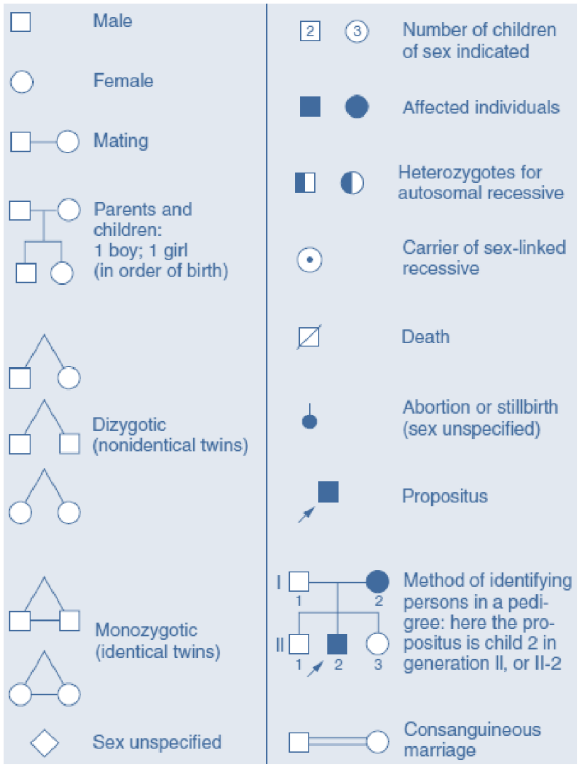
MENDEL´s BREEDING EXPERIMENTS

Phenotypic pea traits in Mendel´s breeding experiments		F_1 results the first Filial generation	F_2 results the second Filial gen. results from F_1xF_1 cross)	F_2 ratio
	round / wrinkled seeds	all round	5474 round 1850 wrinkled	2.96 : 1
	yellow / green seed interior	all yellow	6022 yellow 2001 green	3.01 : 1
	purple / white flower	all purple	705 purple 224 white	3.15 : 1
	smooth / ridged ripe pod	all smooth	882 smooth 299 ridged	2.95 : 1
	green / yellow unripe pod	all green	428 green 152 yellow	2.82 : 1
	axial / terminal flowers	all axial	651 axial 207 terminal	3.14 : 1
	tall / short (dwarf) stem	all tall	787 tall 277 short	2.84 : 1

INHERITANCE – PEDIGREES, TERMINOLOGY AND NOTATION

● A member of a family who first comes to the attention is called the **propositus**. Usually the phenotype of the propositus is exceptional in some way (for example, the propositus might suffer from some type of disorder). The investigator then traces the history of the phenotype in the propositus back through the history of the family and draws a family tree, or pedigree, by using the standard symbols.

● Many variant phenotypes of humans are determined by the alleles of single autosomal genes (like Mendel´s pea phenotypes). The patterns in the pedigree have to be interpreted differently, depending on whether one of the contrasting phenotypes is a rare disorder or whether both phenotypes of a pair are common morphs of a polymorphism. Most pedigrees are drawn up for medical reasons and hence inherently concern medical disorders that are, (almost) by definition, rare.



PEDIGREE ANALYSIS OF AUTOSOMAL RECESSIVE TRAITS

The corresponding unaffected phenotype must be determined by the corresponding dominant allele. For example, the human disease phenylketonuria (PKU) is inherited in a simple Mendelian manner as a recessive phenotype, with PKU determined by the allele *p* and the normal condition by P. Therefore, sufferers from this disease are of genotype *p/p,* and people who do not have the disease are either *P/P or P/p.* What patterns in a pedigree would reveal such an inheritance?

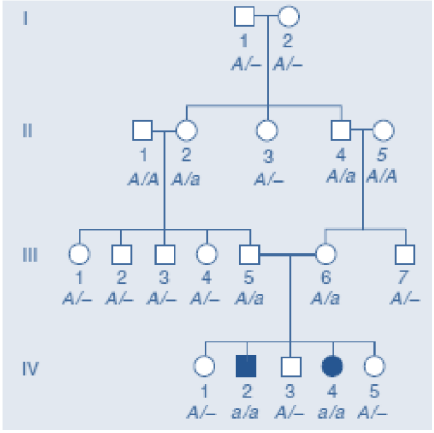
●The two key points are that (1) generally the disease appears in the progeny of unaffected parents and (2) the affected progeny include both males and females. If it is known that both male and female progeny are affected, a reasonbale assumption is a simple Mendelian inheritance of a gene on an autosome, rather than a gene on a sex chromosome. The following typical pedigree illustrates the key point that affected children are born to unaffected parents

● Both parents must be heterozygotes, say *A/a;* both must have an a allele because each contributed an a allele to each affected child, and both must have an *A* allele because they are normal. We can identify the genotypes of the children (in the order shown) as *A/, a/a, a/a, and A/*

● The pedigrees of autosomal recessive traits tend to look rather “bare”, with few black symbols. A recessive condition shows up in groups of affected siblings, and the people in earlier and later generations tend not to be affected. To understand why this is so, it is important to have some understanding of the genetic structure of populations underlying such rare conditions. By definition, if the condition is rare, most people do not carry the abnormal allele

● The formation of an affected person usually depends on the chance union of unrelated heterozygotes. Inbreeding (mating between relatives) increases the chance that two heterozygotes will mate, like cousins in the pedigree. An ancestor who is a Heterozygote may produce many descendants who also are heterozygotes.

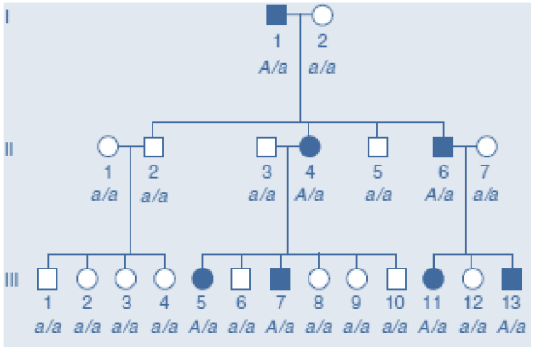
● In general: the smaller the population, the higher the probability that mating individuals are relatives.



PEDIGREE ANALYSIS OF AUTOSOMAL DOMINANT TRAITS

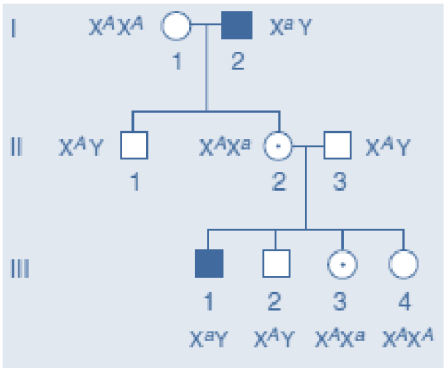
Here the normal allele is recessive, and the abnormal allele is dominant. It may seem paradoxical that a rare disorder can be dominant, but remember that dominance and recessiveness are simply properties of how alleles act and are not defined in terms of how common they are in the population.

● In pedigree analysis, the main clues for identifying an autosomal dominant trait with Mendelian inheritance are that the phenotype tends to appear in every generation of the pedigree and that affected fathers and mothers transmit the phenotype to both sons and daughters.
Abnormal alleles can also arise *de novo* by the process of mutation.



PEDIGREE ANALYSIS OF X-LINKED RECESSIVE TRAITS

Typical features in pedigrees



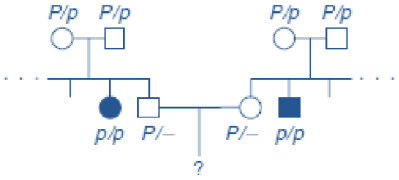
● Many more males than females show the rare phenotype under study. This is because of the product law: a female will show the phenotype only if both her mother and her father bear the allele (for example, XA Xa Xa Y), whereas a male can show the phenotype when only the mother carries the allele. If the recessive allele is very rare, almost all persons showing the phenotype are male.

● None of the offspring of an affected male show the phenotype, but all his daughters are “carriers,” who bear the recessive allele masked in the heterozygous condition. Half the sons of these carrier daughters show the phenotype.

● None of the sons of an affected male show the phenotype under study, nor will they pass the condition to their offspring. The reason behind this lack of male-to-male transmission is that a son obtains his Y chromosome from his father, so he cannot normally inherit the father’s X chromosome, too.

An example problem

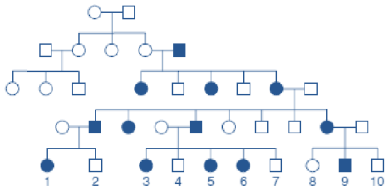
Phenylketonuria (PKU) is a human hereditary disease resulting from the inability of the body to process the chemical phenylalanine, which is contained in the protein that we eat. PKU is manifested in early infancy and, if it remains untreated, generally
leads to mental retardation. PKU is caused by a recessive allele with simple Mendelian inheritance. A couple intends to have children but consults a genetic counselor because the man has a sister with PKU and the woman has a brother with PKU. There are no other known cases in their families. They ask the genetic counselor to determine the probability that their first child will have PKU. What is this probability?



● The only way in which the man and woman can have a PKU child is if both of them are heterozygotes (it is obvious that they themselves do not have the disease). Both the grandparental matings are simple Mendelian monohybrid crosses expected to produce progeny in the following proportions: normal ¾ and PKU ¼. It is known that the man and the woman are normal, so the probability of their being a heterozygote is 2/3, because within the *P/–* class, 2/3 are *P/p* and , *1/3* are *P/P.* The probability of both the man and the woman being heterozygotes is 2/3 x 2/3 = 4/9. If they are both heterozygous, then one-quarter of their children would have PKU, so the probability that their first child will have PKU is ¼ and the probability of their
being heterozygous and of their first child having PKU is 4/9 x ¼ = 4/36 = 1/9, which is the answer.

THIS SET OF PROBLEMS WAS HOME-WORK 2a IN GAME

1. A rare human disease afflicted a family as shown in the pedigree.



- What is the most likely mode of inheritance?
- What would be the outcomes of the cousin marriages 1 x 9, 1 x 4, 2 x 3, 2 x 8 ?

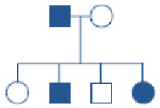
The most likely mode of inheritance is **X-linked dominant**. Dominant, because it appears in every generation. X-linked because fathers do not transmit it to their sons. If it were autosomal dominant, father-to-son transmission would be common. Autosomal recessive is improbable. Note the marriages between affected members of the family and unaffected outsiders. If the condition were autosomal recessive, the only way in which these marriages could have affected offspring is if each person marrying into the family were a heterozygote; then the matings would be *a/a (affected) A/a (unaffected).* However, it is stated that the disease is rare. In such a case, it is highly unlikely that heterozygotes would be so common. X-linked recessive inheritance is impossible, because a mating of an affected woman with a normal man could not produce affected daughters.

Below ***A*** represent the disease-causing allele and ***a*** represent the normal allele.

1x9 Number 1 must be heterozygous *A/a because* she must have obtained *a* from her normal mother. Number 9 must be *A/Y.* The mating is thus *A/a* x *A/Y .* All the daugters will be affected because they inevitably get *A* at least from their father, and half of the sons will be affected because they get Y-chromosome from their father (i.e. they do not inherit the disease causing allele from their father) and half of mother´s X-chromosomes carry *A* and half carry *a.*
1x4 The mating must be *A/a* x *a/Y* **.** Half of children are expected to be affected.
2x3 The mating must be *a/Y* x *A/a ,* like 1x4.
2x8 The mating must be *a/Y* x *a/a ,* all children normal.

THIS SET OF PROBLEMS WAS HOME-WORK 2a IN GAME

2. In the accompanying pedigree, the black symbols represent individuals with a very rare blood disease.



If you had no other information to go on, would you think it most likely that the disease was dominant or recessive? Give your reasons.

Because it is given that the disease is very rare, it is not a good assumption that a heterozygous female and a homozygous male happen to become a couple – and this would be the only case from which affected children would appear. Thus the reasonable explanation is: **dominant.**

3. The ability to taste the chemical phenylthiocarbamide is an autosomal dominant phenotype, and the inability to taste it is recessive. If a taster woman with a nontaster father marries a taster man who in a previous marriage had a nontaster daughter, what is the probability that their first child will be ● A nontaster girl ● A taster girl ● A taster boy ● What is the probability that their first two children will be tasters of either sex.

A nontaster girl. Let *B* and *b* be te alleles for taster and nontaster, resp. The woman´s father is a nontaster and must be *bb.* The woman is a taster, and must be *Bb.* The man is taster and since he has a nontaster daughter, he must have the allele *b* to pass down to his daughter. Hence, the man is also heterozygote *Bb.* Of all four possible outcomes, *BB, Bb, bB, bb* (genotypes of children) only one is *bb,* nontaster, hence the probability of getting a taster child is ¼. The probability of child being a girl is ½. ¼ x ½ = 1/8.
A taster girl = a taster boy. ¾ x ½ = 3/8 (¾ because 3 out of 4 possible children genotypes result in tasters).
First two children tasters of either sex. ¾ x ¾ = 9/16.

THIS SET OF PROBLEMS WAS HOME-WORK 2a IN GAME

4. Bill and Mary are contemplating having children, but Bill´s brother has galactosemia (an autosomal recessive disease) and Mary's great-grandmother also had galactosemia. Mary has a sister who has three children, none of whom have galactosemia.
● What is the probability that Bill´s and Mary´s s first child will have galactosemia?

Let *A* and *a* be the normal and disease alleles, respectively .
Mary´s great grandmother is *aa*, since it is given that she had the disease.
Since galactosemia is rare we can assume that she married a *AA* man.
Hence, Mary´s grandparent is heterozygote *Aa.*
We can again assume that the grandparent married a *AA* (because the disease is rare).
Thus, Mary´s parent could be either *Aa* or *aa,* with probability ½ each.
Similarly, given that Mary´s parent has genotype *Aa,* the probability of Mary being heterozygote is again ½ .
The other parent is, again, *AA.*
The probabilities for Mary´s genotype are, therefore Pr(*Aa*) = ½ x ½ = ¼ Pr(*AA*) = ½ x 1 + ½ x ½ = ¾
The fact that Mary has a sister with 3 normal children does not provide any useful hints for this problem.

Bill´s brother has galactosemia, i.e. Bill´s brother is *aa* and nothing is given about their parents => they are both healthy, but carriers (otherwise it would be impossible to get *aa* child (=Bill).
Thus, both Bill´s parents are *Aa*, which results in the following probabilities for Bill´s genotype:
Pr (*AA*) = 1/3 Pr(*Aa*) = 2/3

Therefore, Pr(Bill´s and Mary´s first child will have galactosemia)
= Pr(Bill is *Aa) x* Pr (May is *Aa*) x Pr(child is *aa* given both parents are *Aa*) = ¼ x 2/3 x ¼ = 1/24

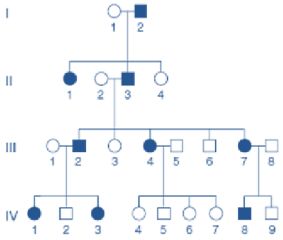
THIS SET OF PROBLEMS WAS HOME-WORK 2a IN GAME

5. Suppose that a husband and wife are both heterozygous for a recessive allele for albinism. ● If they have dizygotic (two-egg) twins, what is the probability that both the twins will have the same phenotype for pigmentation?

Pr (normal) = ¾ Pr(albino) = ¼
Pr(twins have the same phenotype for pigmentation) = Pr(both normal) + Pr(both albino) = ¾ x ¾ + ¼ x ¼ = 5/8.

6. The pedigree is for a rare, but relatively mild, hereditary disorder.

- Is the disorder inherited as a recessive or a dominant phenotype?
- Give genotypes for as many individuals in the pedigree as possible. Invent your own defined allele symbols.
- Consider the four unaffected children of parents III-4 and III-5. In all four-child progenies from parents of these genotypes, what proportion is expected to contain all unaffected children?



The disorder is rare, so for a given carrier his/her mate is non-carrier (the reasonable assumption). There are too many affected individuals, and in all generations in the pedigree to lead to a conclusion of a recessive inheritance. Answer: dominant.

I *aa aA*
II *aA aa aA aa*
III *aa aA aa aA aa aa aA aa*
IV *aA aa aA aa aa aa aa aA aa*

aA x aa → Prob ½ *aa* (healthy) prob ½ *aA* (affected) all 4 healthy ½ x ½ x ½ x ½ = 1/16

THIS SET OF PROBLEMS WAS HOME-WORK 2a IN GAME

7. Tay-Sachs disease is a rare human disease in which toxic substances accumulate in nerve cells. The recessive allele responsible for the disease is inherited in a simple Mendelian manner. A woman is planning to marry her first cousin, but the couple discovers that their shared grandfather's sister died in infancy of Tay-Sachs disease.

●What is the probability that the cousins' first child will have Tay-Sachs disease, assuming that all people who marry into the family are homozygous normal?

G normal allele, *g* disease allele. Grandfather´s sister must be *gg* and great-grandparents must both have been *Gg.* The probabilities for woman´s grandfather: Pr(*Gg*) = 2/3 Pr(*gg*) = 1/3

Assuming that this grandfather married a homozygous normal, we get these mating tables

	G	*g*
G	*GG*	*Gg*
G	*GG*	*Gg*

	g	*g*
G	*Gg*	*Gg*
G	*Gg*	*Gg*

with probabilities 2/3 1/3

From these tables we obtain the probability that the woman´s (and the man´s) parent is heterozygous: 2/3 x 1 / 2 = 1/3 homozygous 1/3 x 1 + 2/3 x 1/2 = 2/3
Whatever the parent is, he/she married a homozygous normal (rare disease, reasonable assumption), so the probability that the woman is heterozygous is 1/3 x 1/2 = 1/6 since her parent has to be heterozygous himself/herself - 1/3 - and pass down the allele accordingly - ½. Similarly the probability that the woman is homozygous normal is 2/3 x 1 + 1/3 x ½ = 5/6. Her cousin has the very same probabilities. The probability that their first child will have Tay-Sach´s disease is therefore

,

1/6 x 1/6 x 1/4 = 1/144

THIS SET OF PROBLEMS WAS HOME-WORK 2a IN GAME

8. A man's grandfather has galactosemia. This is a rare autosomal recessive disease caused by inability to process galactose, leading to muscle, nerve, and kidney malfunction. The man married a woman whose sister had galactosemia. The woman is now pregnant with their first child.
   - Draw the pedigree as described.
   - What is the probability that this child will have galactosemia?
   - If the first child does have galactosemia, what is the probability a second child will have it.

The grandfather must have been homozygous for the disease allele.
Thus the man´s parent is heterozygous.
The probability that the man carries the allele is ½
Woman´s sister had galactosemia => both the parents of the woman must be disease allele carriers and since neither of them have galactosemia, they are heterozygotes.
Therefore, the probability that the woman carries the allele is 2/3.

Pr(first child will have galactosemia) = ½ x 2/3 x ¼ = 1/12

First child has galactosemia => both parents must be heterozygotes (carriers) => ¼
(because ¼ of progeny from het x het mating are recessive homozygotes, affected)

THIS SET OF PROBLEMS WAS HOME-WORK 2a IN GAME

9. In humans, color vision depends on genes encoding three pigments. The *R* (red pigment) and *G* (green pigment) genes are on the X chromosome, whereas the *B* (blue pigment) gene is autosomal. A mutation in any one of these genes can cause colorblindness. Suppose that a colorblind man married a woman with normal color vision. All their sons were colorblind, and all their daughters were normal.
- Specify the genotypes of both parents and all possible children, explaining your reasoning.

It is clear that the disease is X-linked The sons have an X-chromosome with the allele for color-blindedness which they received from their mother. Father X_c Y, mother X_c X and son like his father and daughter like her mother.

FROM GAME

THE BASIC MODEL IN POPULATION GENETICS: HARDY-WEINBERG

● Mendel´s laws of inheritance were re-discovered in 1900 but it was not understood how genetic variation behaves from generation to generation. For example, it was thought that a dominant allele should increase in frequency .

● In 1908, famous British mathematician, G.H. Hardy wrote: *To the Editor of Science: I am reluctant to intrude in a discussion concerning matters of which I have no expert knowledge, and I should have expected the very simple point which I wish to make to have been familiar to biologists. However, some remarks of Mr. Udny Yule, to which Mr. R. C. Punnett has called my attention, suggest that it may still be worth making... Suppose that Aa is a pair of Mendelian characters, A being dominant, and that in any given generation the number of pure dominants (AA), heterozygotes (Aa), and pure recessives (aa) are as p:2q:r. Finally, suppose that the numbers are fairly large, so that mating may be regarded as random, that the sexes are evenly distributed among the three varieties, and that all are equally fertile. A little mathematics of the multiplication-table type is enough to show that in the next generation the numbers will be as* $(p+q)^2:2(p+q)(q+r):(q+r)^2$*, or as* $p_1:2q_1:r_1$*, say. The interesting question is — in what circumstances will this distribution be the same as that in the generation before? It is easy to see that the condition for this is* $q^2 = pr$*. And since* $q_1^2 = p_1 r_1$*, whatever the values of p, q, and r may be, the distribution will in any case continue unchanged after the second generation*

● In 1908, German physicist Wilhelm Weinberg published the same result, independently.

HARDY-WEINBERG ”EQUILIBRIUM”, ASSUMPTIONS

● Summary of assumptions:
- The organism is diploid and reproduces sexually.
- Generations are nonoverlapping.
- The gene under consideration has two alleles, ***A*** and ***a.***
- The allele frequencies, ***p*** and ***q***, in the population, consisting of the individuals (genotypes), ***AA, Aa, aa*** are identical in males and females.
- Mating is random.
- Population size is very large (infinite).
- Migration is negligible.
- Mutation at the gene locus we consider is so rare that can be ignored (i.e. ***A*** mutating to ***a***, or ***a*** to ***A***).
- Natural selection does not affect the alleles under consideration.

● The assumption of infinite population means that random (stochastic) events can be ignored. Negligible migration means that, if there is another population with different allele frequencies, change of individuals (=migration) does not disturb the situation and violate the assumption of a closed system. Natural selection here means that ***A*** and ***a,*** or ***AA, Aa, aa*** perform equally well in reproduction, there are no *fitness* differences.

●**These assumptions summarize the Hardy-Weinberg model.**

GENOTYPE FREQUENCIES – ALLELE FREQUENCIES

- In Hardy-Weinberg model the relation between the *allele* frequencies, p and q ($p + q = 1$), and the *genotype* frequencies is given by

$$AA: p^2 \qquad Aa: 2pq \qquad aa: q^2,$$

- The formation of one generation from the previous generation as an outcome of repeated and independent trials (assuming random mating the choices of male gamete and female gamete are independent trials):

  pairs of gametes (carrying the alleles A and a), AA, Aa and aa, are expected in proportions given by

$$(p + q)^2 = p^2 + 2pq + q^2$$

			Male gametes	
		allele	A	a
		frequency	p	q
Female gametes	A	p	AA p^2	Aa pq
	a	q	aA qp	aa q^2

MATING TYPE FREQUENCIES => NEXT GENERATION (= THE PROGENY POP.) GENOTYPE AND ALLELE FREQUENCIES

Frequency of zygotes (progeny)

Mating	Frequency of mating	AA	Aa	aa
AA x AA	P^2	1	0	0
AA x Aa	2PQ	½	½	0
AA x aa	2PR	0	1	0
Aa x Aa	Q^2	¼	½	¼
Aa x aa	2QR	0	½	½
aa x aa	R^2	0	0	1
	Totals (next gen.)	P´	Q´	R´

$P' = P^2 + (2PQ)/2 + Q^2/4 = (P + Q/2)^2 =$
$Q' = (2PQ)/2 + 2PR + Q^2/2 + (2QR)/2 = 2(P + Q/2)(R + Q/2) =$
$R' = Q^2/4 + (2QR)/2 + R^2 = (R + Q/2)^2 =$

p^2
$2pq$
q^2

So, random mating of genotypes - random union of gametes - HWE.

HWE

● With two alleles of gene, there are six possible types of matings (the left column in the table)

● When mating is random, the matings take place in proportion to the genotypic frequencies in the population, and the types of mating pairs are given by successive terms in the expansion of

$$(P\,AA + Q\,Aa + R\,aa)^2$$

● The proportion of AA x AA matings is P x P = P^2 and the proportion of AA x Aa matings is 2 x P x Q because the mating can be between either an AA female and an Aa male (P x Q) or Aa female and AA male (Q + P). The frequencies of these and the other types of matings are given in the second column.

● **The next generation** (zygotes from which progeny follow): ***Mendel´s law of segregation is taken into account.***

● Aa heterozygote produces and equal number of A-bearing and a-bearing gametes. AA and aa homozygotes produce only A and a gametes, respectively.

● The mating AA x aa produces all Aa zygotes, the mating AA x Aa produces ½ AA and ½ Aa zygotes, Aa and Aa produces ¼ AA, ½ Aa, ¼ aa, etc.

HWE

● Why a model with so many restrictive assumptions? Are all these assumptions likely to be met in actual populations?

● HWE is not meant to be an exact description of any actual population, although actual populations often exhibit genotype frequencies predicted by it.

● HWE provides a **null model,** a prediction based on a simplified or idealized situation where no biological processes are acting and genotype frequencies are the result of random combination.

● Actual populations can be compared with this null model to test hypotheses about the evolutionary forces acting on allele and genotype frequencies.

● The important point and the original motivation for Hardy and Weinberg was to show that the process of particulate inheritance itself does not cause any
changes in allele frequencies across generations.

● Thus, changes in allele frequency or departures from HWE expected genotype frequencies must be caused by processes that alter the outcome of basic inheritance.

THIS SET OF PROBLEMS WAS HOME-WORK 2b IN GAME

1. The table below shows the observed numbers of *AA, Aa* and *aa* genotypes in samples of size 100 from four populations, 1-4. For which samples can the hypothesis of Hardy-Weinberg proportions be rejected?

population	*AA*	*Aa*	*aa*
1	8	53	39
2	9	61	30
3	13	58	29
4	18	35	47

Observed allele frequencies and **expected** (=HW) **genotype frequencies**:
p for *A* and *q* for *a*

	p^2	$2pq$	q^2
In the sample 1:			
p = (2 x 8 + 53) / 200 = 0.35	0.35 x 0.35 x 100 = 11.9	2 x 0.35 x 0.65 x 100 = 45.2	0.65 x 0.65 x 100 = 42.9
q = 1 – p = 0.65			

The statistical question is, whether the observed numbers, 8 53 39 and expected numbers 11.9 45.2 42.9 match.

THIS SET OF PROBLEMS WAS HOME-WORK 2b IN GAME

- A conventional statistical test assessing quantitatively the closeness of fit is the **chi-square test**:

Chi-square = Σ [(*observed* – *expected*)^2 / *expected*] and the **degrees of freedom** (*df*) = number of classes of data minus number of parameters estimated from the data minus 1.

chi-square is 2.98 and *df* = 3 – 1 – 1 = 1 Three classes, one parameter (*p*) must be estimated. => P = 0.08 => HW-proportions ok

- Similar calculations for other samples

Sample 2 *chi-square* = 7.63 P = 0.006 => HW-proportions do not hold (null hypothesis rejection at ”1% significance level”)

Sample 3 *chi-square* = 3.63 P = 0.057 => HW-proportions ok (if the P value would have been smaller than 0.05, then the null hypothesis would have been rejected at ”5% significance level”)

Sample 4 *chi-square* = 5.56 P = 0.018 => like sample 2

THIS SET OF PROBLEMS WAS HOME-WORK 2b IN GAME

2. For a trait due to a rare X-linked recessive allele, show that the frequency of heterozygous carrier females is approximately equal to two times the affected males. Calculate the frequencies for an X-linked recessive allele with an allele frequency of 0.1

The recessive allele frequency is marked by *q*. As males have only one X-chromosome, then in males the the allele frequencies and genotype frequencies are the same (males are haploid as regards X-chromosomal genes). The frequency of affected males is thus *q*.

The frequency of heterozygous carrier females equals $2pq = 2(1 - q)q = 2q - 2q^2 \approx 2q$ because when the recessive allele is rare (q is small), then $q^2 \approx 0$.

The frequency of carrier females is thus approximately two times the the frequency of affected males.

For $q = 0.1$, the frequency of heterozygous carrier females is 2 x 0.1 x 0.9 = 0.18

THIS SET OF PROBLEMS WAS HOME-WORK 2b IN GAME

3. The gene *CCR5* encodes a protein co-receptor used by the AIDS virus for entry into certain white blood cells. Many populations are polymorphic for a deletion of part of the coding sequence that results in an inactive protein. This polymorphism was originally discovered among persons infected with the virus who had remained free of the AIDS disease for at least 10 years. The protective effect of the deletion, denoted *CCR5 Δ* , is at least a factor of two. In one study of 338 individuals from one human population the observed numbers of the genotypes were as follows: 265 *CCR5 / CCR5,* 66 *CCR5 / CCR5 Δ* , 7 *CCR5 Δ / CCR5 Δ* .

Is there any reason to reject the hypothesis of Hardy-Weinberg proportions for this gene?

Allele frequency and expected genotype frequency calculations as in problem 1.

$p = 0.88$ and $q = 0.12$ => expected numbers of genotypes are 262.9, 70.4, 4.7 and compared with observed numbers (265, 66, 7) => *chi-square* 1.42, *df* = 1, probabality 0.25 => no reason to reject the hypothesis of HW-proportions.

THIS SET OF PROBLEMS WAS HOME-WORK 2b IN GAME

4. In a population sample of 1617 individuals the numbers of A, B, O and AB blood types observed were 724, 110, 763 and 20. The best estimates of allele frequencies are: 0.26 for the allele I^A , 0.04 for the allele I^B and 0.69 for the allele I^O . The genotypes $I^A I^A$ and $I^A I^O$ product phenotype A, genotypes $I^B I^B$ and $I^B I^O$ product phenotype B, $I^O I^O$ phenotype O and $I^A I^B$ phenotype AB **.**

Calculate the expected numbers of the four phenotypes (the blood types A, B, O, AB) .
Are the blood types in HW-proportions?

Extension of the two-allele case, $(p + q)^2 = p^2 + 2pq + q^2$,

into three alleles $(p + q + r)^2 = p^2 + q^2 + r^2 + 2pq + 2pr + 2qr$ (*)

Let p the fr of allele I^A , q the fr of allele I^B and r the fr of allele I^O .

Then the expected numbers of bloodtypes are 710.7, 94.8, 776.1, 35.4 (see the question: it tells the relationships between genotypes and phenotypes, and combine this information with (*), then you get these numbers).

Chi-square is 9.61 , *df* = 1 (there are 4 classes, two parameters estimated, minus one) corresponding probability is 0.002.

HW proportions thus do not hold .

THIS SET OF PROBLEMS WAS ASSIGNMENT 2b IN GAME12-COURSE

5. Colorblindness results from a sex-linked recessive allele. One in every ten males is color-blind. Consider a population in HW-proportions:
   - What proportion of all women are color-blind?

Let X^a denote the recessive allele leading to color blindness and X^A the normal allele, frequencies q and p, respectively.
Colorblind females $q^2 = 0.01$. Note that q was given, it is 1/10 (allele frequencies = genotype frequencies in males as they are haploid as regards X-chromosomal genes.

- By what factor is color blindness more common in men? Or, how many color-blind men are there for each color-blind woman)? Answer: 10:1

- In what proportion of marriages would color blindness affect half the children of each sex?

Regardless of the sex, half of the children will be colorblind when a colorblind man has married a woman who is heterozygous. Probabality for this kind of marriage is $q \times 2pq = 0.018$

- In what proportion of marriages would all children be normal?

All children will be normal when colorblind man (X^a Y) marries a woman who is homozygous $X^A X^A$ and when man is not colorblind (X^A Y) and woman is $X^A X^A$:

0.1 x 0.81 + 0.9 x 0.81 = 0.81

THIS SET OF PROBLEMS WAS ASSIGNMENT 2b IN GAME12-COURSE

6. Genotype frequencies and HW in forensics – DNA profiling.
A crime has been committed. Left at the crime scene was a biological sample that law-enforcement authorities use to obtain a multilocus genotype or DNA profile. A suspect in the crime has been identified and subpoenaed to provide a tissue sample for DNA profiling. The DNA profiles from the suspect and from the crime scene are identical. Should we conclude that the suspect left the biological sample found at the crime scene ? The DNA-profile is this:

DNA-profiling for individual identification is commonly performed by using STR-loci (short tandem repeats, microsatellites). These have very large number of alleles. The alleles are various versions of DNA-repeats. In the example 17, 18 means that at the locus D3S1358 the individual is heterozygous for repeat counts 17 and 18. The locus is in 3rd chromosome (= D3, S1358 depicts its detailed location).

Locus	**D3S1358**	**D21S11**	**D18S51**
Genotype	17, 18	29, 30	18, 18

To answer this question HW prediction of the expected frequency of the DNA profile or genotype is one elementary step. Just because two DNA profiles match, there is not necessarily strong evidence that the individual who left the evidence DNA and the suspect are the same person. It is possible that there are actually two or more people with identical DNA profiles. HW and Mendel's second law of inheritance will serve as the bases to estimate just how frequently a given DNA profile should be observed. Then it is possible to determine whether two unrelated individuals sharing an identical DNA profile is a likely occurrence.

This example and exercise is taken from:
Hamilton, Population genetics, 2009, Wiley-Blackwell

THIS SET OF PROBLEMS WAS ASSIGNMENT 2b IN GAME12-COURSE

6. continues…
Allele frequencies for 9 STR loci used in forensic cases (FBI data), based on a sample of 196 US white citizens, sampled randomly with respect to geographic location.

D3S1358		vWA		D21S11		D18S51		D13S317		FGA		D8S1179		D5S818		D7S820	
Allele	Freq	Allele	Freq	Allele	Freq	Allele	Freq	Allele	Freq	Allele	Freq	Allele	Freq	Allele	Freq	Allele	Freq
12	0.0000	13	0.0051	27	0.0459	<11	0.0128	8	0.0995	18	0.0306	<9	0.0179	9	0.0308	6	0.0025
13	0.0025	14	0.1020	28	0.1658	11	0.0128	9	0.0765	19	0.0561	9	0.1020	10	0.0487	7	0.0172
14	0.1404	15	0.1122	29	0.1811	12	0.1276	10	0.0510	20	0.1454	10	0.1020	11	0.4103	8	0.1626
15	0.2463	16	0.2015	30	0.2321	13	0.1224	11	0.3189	20.2	0.0026	11	0.0587	12	0.3538	9	0.1478
16	0.2315	17	0.2628	30.2	0.0383	14	0.1735	12	0.3087	21	0.1735	12	0.1454	13	0.1462	10	0.2906
17	0.2118	18	0.2219	31	0.0714	15	0.1276	13	0.1097	22	0.1888	13	0.3393	14	0.0077	11	0.2020
18	0.1626	19	0.0842	31.2	0.0995	16	0.1071	14	0.0357	22.2	0.0102	14	0.2015	15	0.0026	12	0.1404
19	0.0049	20	0.0102	32	0.0153	17	0.1556			23	0.1582	15	0.1097			13	0.0296
				32.2	0.1122	18	0.0918			24	0.1378	16	0.0128			14	0.0074
				33.2	0.0306	19	0.0357			25	0.0689	17	0.0026				
				35.2	0.0026	20	0.0255			26	0.0179						
						21	0.0051			27	0.0102						
						22	0.0026										

At the locus D3S1358, we see from this background reference table that the 17-repeat allele has a frequency of 0.2118 and the 18-repeat allele a frequency of 0.1626. Using HW, the 17, 18 genotype has an expected frequency of 2(0.2118)(0.1626) = 0.0689 or 6.89%. For the two other loci in the DNA profile the expected frequencies are 2(0.1811)(0.2321) = 0.0841 or 8.41%, and $(0.0918)^2$ = 0.0084 or 0.84%.

THIS SET OF PROBLEMS WAS ASSIGNMENT 2b IN GAME12-COURSE

6. continues…

The genotype for each locus has thus a relatively large chance of being observed in the population. For example, about 1 in 119 are expected to be homozygous for the 18-repeat allele at locus D18S51. Therefore, a match between evidence and suspect DNA profiles homozygous for the 18 repeat at that locus would not be strong evidence.

Combining the information from all three loci:

The expected frequency under HW, and under the assumption that each locus is independent by Mendel's second law (they are on different chromosomes). The expected frequency of the three locus genotype (sometimes called the probability of identity) is then 0.0689 × 0.0841 × 0.0084 = 0.000049 or 0.0049%.

Another way to express this probability is as an **odds ratio,** or the reciprocal of the probability (an approximation that holds when the probability is very small).

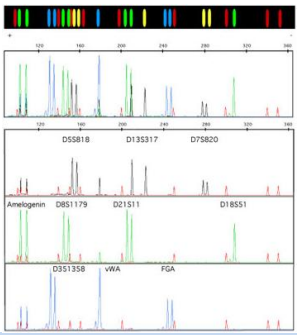
Here the odds ratio is 1/0.000049 = 20,408, meaning that we would expect to observe the three-locus DNA profile once in 20,408 white US citizens.

Forensic DNA profiles use 10–13 loci to estimate expected genotype frequencies.

The 10-locus genotype for the same individual:

D3S1358	17, 18
vWA	17, 17
FGA	24, 25
Amelogenin	X, Y
D8S1179	13, 14
D21S11	29, 30
D18S51	18, 18
D5S818	12, 13
D13S317	9, 12
D7S820	11, 12

Calculate the expected genotype frequency and odds ratio for the 10-locus genotype. Would a match between a crime scene sample and a suspect be convincing evidence?



THIS SET OF PROBLEMS WAS ASSIGNMENT 2b IN GAME12-COURSE

The main part of this question involves "practical education" about DNA-forensics and the final question is: continue the computation as already shown, take the frequencies from the table (amelogenin 0.5, the gene marker which is used for man/female distinction) .

All loci are in different chromosomes => multiplications of the genotype frequencies.

Expected 10-locus genotype frequency will be

$2 \times (0.2118 \times 0.1626) \times (0.2628)^2 \times \ldots\ldots\text{etc} = 1.4 \times 10^{-12}$ (maybe something else than 1.4, the order of magnitude is correct)

Odds ratio 1 / prob = 8×10^{11} meaning that we would expect to observe the 10-locus DNA-genotype once in this many people , pointing out that the evidence is convincing.

THIS SET OF PROBLEMS WAS ASSIGNMENT 2b IN GAME12-COURSE

7. Suppose there are two populations that have genotype frequencies

	AA	*Aa*	*aa*
Pop 1	0.64	0.32	0.04
Pop 2	0.09	0.42	0.49

If a researcher draws a sample, thinking it is coming from a single population, but it is actually composed of individuals two-thirds of whom came from population 1, and one-third from population 2,

● If these individuals are simply collected together, but have not really interbred, what will the genotype frequencies in the sample expected to be?

2/3 x 0.64 + 1/3 x 0.09 = 0.457 (*AA*)
2/3 x 0.32 + 1/3 x 0.42 = 0.353 (*Aa*) 2/3 x 0.04 + 1/3 x 0.49 = 0.190

● What will the allele frequencies be expected to be in that sample?

2/3 x 0.8 + 1/3 x 0.3 = 0.633 (*A*, 0.8 in pop 1 and 0.3 in pop 2) 1 – 0.633 = 0.367

● If we mistakenly assume that the sample is from a random-mating population, and use the sample allele frequency, what proportion of heterozygotes will we expect to see?

2 x 0.633 x 0.367 = 0.46

THIS SET OF PROBLEMS WAS ASSIGNMENT 2b IN GAME12-COURSE

8. A locus has three alleles, *B´, B, and b.*
B´ is completely dominant to *B,* and both of these are completely dominant to *b.*
● What are the frequencies of the three alleles in a random-mating population which has these phenotype frequencies: 50% *B´-, 30% B-, and 20% bb ?*

Let the allele frequencies be p for B' , q for B and r for b.

See the solution of question 4 for HW-proportions in a 3-allele case.
Assuming that HW-proportions hold gives tools for allele frequency estimation.

The frequency of bb is $r^2 = 0.2$ $r = 0.45$ (the frequency of b – allele).

Frequency $BB + Bb$ is $q^2 + 2qr = 0.3$ $q = 0.26$

$p = 1 – q - r = 0.29$ (the frequency of B' - allele)

ESTIMATING ALLELE FREQUENCIES

● Consider a **sample** of *n* diploid individuals drawn from a random-mating population and the problem of estimating the allele frequency p_A in the **population.**

● Suppose that we sampled 100 individuals, and found 49 *AA*, 26 *Aa*, and 25 *aa*

■ We could estimate the allele frequency in the population by simply taking the allele frequency in the sample. This gives p_A = (98 + 26)=200 = 0.62
■ We could also consider that we expect the proportion of AA individuals in the sample to be (on the average) the same as the population genotype frequency $p^2_{A.}$ So we could take the observed frequency of AA, 0.49, and take its square root to get an estimate of the gene frequency, 0.7.
■ We could also take the square root of the observed frequency (0.25) of aa, which gives an estimate of 0.5 for the frequency of a, and hence 0.5 for the frequency of A.

■ Now we have three different estimates (0.5, 0.62, and 0.7) for the same quantity. *All these methods will give an allele frequency close to that in the population, if the sample size is large. But which estimate is to be preferred when it is not?*
■ Posing the problem as a statistical one, and using some standard statistical approach (minimum variance unbiased estimates, minimum mean square error methods, Bayesian and empirical Bayesian approaches).

● We use here maximum likelihood (ML) method.

ML IN ESTIMATING ALLELE FREQUENCIES

● ML in general:
Suppose that we want to estimate a parameter, Θ , and are given some data. If we have a probabilistic model for the generation of the data, we could compute for a given value of Θ, the probability Prob (*Data* | Θ) that the observed set of data would have arisen. (This is not to be confused with Prob (Θ | Data), which would be the probability of a particular value of Θ , given the data.)
● The method of maximum likelihood is to vary Θ until we find that value which maximizes Prob(Data | Θ), the probability of the data, given Θ. Prob(Data | Θ) is referred to as the likelihood of Θ . Considered as a function of the data, it is a probability. But for a fixed set of data, as a function of Θ , it is called a likelihood.
● ML method has a number of desirable properties. As the sample size increases, the estimate will approach the true value of Θ. For a given sample size (provided it is large), the variance of the estimate Θ of around the true value is less under the ML method than under any other. The estimate is not necessarily unbiased (that is, the average estimate of Θ on repeated sampling may not be exactly Θ), but the amount of bias declines as sample size increases.

● Back to our allele frequency estimation problem:
The data are the numbers of the genotypes observed in the sample. Suppose that these are n_{AA}, n_{Aa}, n_{aa}: The role of Θ is played by the unknown gene frequency *p*. We need to know how to compute Prob(n_{AA}; n_{Aa}; n_{aa} | *p*). We have a sample of *n* individuals, drawn from a population in which the true genotype frequencies are p^2, 2p(1-p), $(1-p)^2$. The probability of the observed numbers n_{AA}; n_{Aa}; n_{aa} is the multinomial probability

$$\text{Prob}\,(n_{AA}, n_{Aa}, n_{aa} \mid p) = \binom{n}{n_{AA}\ n_{Aa}\ n_{aa}} (p^2)^{n_{AA}}\,[2p(1-p)]^{n_{Aa}}\,[(1-p)^2]^{n_{aa}} \qquad \textbf{(3)}$$

● Equation (3) can be rewritten as

$$\text{Prob}\,(n_{AA}, n_{Aa}, n_{aa} \mid p) = C\, p^{2n_{AA} + n_{Aa}} (1 - p)^{2n_{aa} + n_{Aa}} \quad \textbf{(4)}$$

where C incorporates the constant terms and the factorials which depend on the n´s but not on p.

● Varying p to maximize the likelihood is easier by using logarithms:

$$\log_e L = \log_e C + (2n_{AA} + n_{Aa}) \log_e p + (n_{Aa} + 2n_{aa}) \log_e (1 - p) \quad \textbf{(5)}$$

● Plotting $\log_e L$ as a function of p, reaching the maximum, the slope of the curve will be zero. Trying to find the value of p at this point: derivative of (5) and equating it to zero

$$d \log_e L / dp = (2n_{AA} + n_{Aa}) / p - (n_{Aa} + 2n_{aa}) / (1 - p) = 0 \quad \textbf{(6)}$$

● The value of p which solves (6) is

$$p = (2n_{AA} + n_{Aa}) / 2n \quad \textbf{(7)}$$

● This means that 0.62 is the ML estimate in the example above.

TESTING FOR HW-PROPORTIONS

● The common practice is to use *chi-square test*: *observed* number of genotypes in each class (with two alleles there are three genotype classes), *expected* number, etc.

$$\text{chi-square} = \Sigma\ (\text{observed} - \text{expected})^2 / \text{expected}$$

Example 1

● An amino acid (ah) polymorphism at p53 gene: at position 72 either arginine (Arg) or proline (Pro). Let´s call them *Arg* and *Pro* alleles. Among 318 humans in one population: 166 *Arg/Arg,* 120 *Arg/Pro,* 32 *Pro/Pro.* Allele frequencies are thus:

for *Arg*-allele (2 x 166 + 120) / (2 x 318) = 0.71

for *Pro*-allele (2 x 32 + 120) / (2 x 318) = 0.29

● In HW the genotypes should be in proportions $(0.71)^2 = 0.505$, $2 \times 0.71 \times 0.29 = 0.411$ and $(0.29)^2 = 0.084$, i.e. 160.6, 130.8 and 26.6 individuals.
Chi-square: $(166\text{-}160.6)^2 / 160.6 + (120 - 130.8)^2 / 130.8 + (32 - 26.6)^2 / 26.6 = 2.17$
One degree of freedom (df) as there are 3 classes, and one estimated parameter (allele frequency). The corresponding probability value is (see chi-squre table) P = 0.14. The generally agreed cutoff for a significantly low P is 0.05 (goodness of fit considered poor that the model us judged invalid for the data). In this example there are no reason to reject the hypothesis that the genotype frequencies are in HW-proportions.

EXACT TEST FOR HW-PROPORTIONS

● Sample size is too small for a traditional chi-square test,
let the observed numbers of *AA, Aa, aa* in one possible sample be n_{11}, n_{12}, n_{22} , total sample size is $n = n_{11} + n_{12} + n_{22}$ and the observed numbers of alleles *A* and *a* are $n_1 = 2 \times n_{11} + n_{12}$
$n_2 = 2 \times n_{22} + n_{12}$

● Calculation of all possible sample configurations (n_{11}, n_{12}, n_{22}) for a fixed sample size *n* and fixed allele counts n_1 and n_2 .

■ The exact probability of the sample configuration (n_{11}, n_{12}, n_{22}), conditional of the allele counts (n_1 , n_2) is

$$\Pr\{n_{12} | n_1, n_2\} = \{[(n! / n_{11}! n_{12}! n_{22}!)] / [(2n)! / (n_1! n_2!)]\} / 2^{n_{12}} \quad \textbf{(4)}$$

● Once these conditional probabilities have been calculated for all possible values of n_{12} , they are arranged in increasing order, and a cutoff is chosen such thatr the cumulative probability of all outcomes above the cutoff equals 0.05 (or smaller than 0.05).If the observed genotype counts fall below the cutoff, the hypothesis of HW is rejected.

● The exact test is the most common test of significance fro departures from HWE in small samples. In practice, *P* values are calculated using either some standard statistical software package or some web-based calculator (google: ”exact test for Hardy-Weinberg).

EXACT TEST FOR HW-PROPORTIONS

Example 2

● Consider a sample of size 8 diploid individuals with fixed allele counts $n_1 = 8$ and $n_2 = 8$. There are five possible sample configurations (n_{11}, n_{12}, n_{22}):

	Probability (see eq. (4))
(0, 8, 0)	0.0199
(1, 6, 1)	0.2785
(2, 4, 2)	0.5222
(3, 2, 3)	0.1740
(4, 0, 4)	0.0054

● Probabilities/sample configurations in increasing order of pr.

		cumulative probabilities:
(4, 0, 4)	0.0054	0.0054
(0, 8, 0)	0.0199	0.0054 + 0.0199 = 0.0253
(3, 2, 3)	0.1740	0.0253 + 0.1740 = 0.1994
(1, 6, 1)	0.2785	0.1194 + 0.2785 = 0.4779
(2, 4, 2)	0.5222	0.4779 + 0.5221 = 1.0000

● In each row the cumulative probability value corresponds to the *P* value of observing a fit as bad (or worse) than the sample configuration, given in that row. An observed sample configuration (4, 0,4) would lead to rejection of the hypothesis of HW with a significance level of 0.0054, and an observed sample (0,8,0) rejection of HW with the significance level of 0.0253.

EXACT TEST FOR HW

Example 3

● Consider two alleles, three genotypes 66 heterozygotes, 265 and 7 homozygotes.

● Thus $n_1 = 596$ and $n_2 = 80$. There are 41 sample configurations that are compatible with these, have the form $(n_{11}, n_{12}, n_{22}) = (298 - x, 2x, 40 - x)$, where x can assume the values 0,1,2,…., 40. Each of these possible samples has a porbability of occurrence given by equation (4). Here the chi-square is used only as a measure of the magnitude of the deviation, without assuming that the values are actually distributed as X^2. Among the 41 possibilites, 37 yield chi-square values as great or greater than the observed value, and these samples have a cumulative probability of 0.290. This is the exact P value. If chi-square calculated by (3) and P-value from chi-square table: $P = 0.25$.

PERMUTATION TEST FOR HW

Example 2 again: Consider a large number of random permutations of (1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,18), where even numbers represent one allele and odds numbers the other allele. Each successive pair of numbers would then constitute one diploid genotype in the sample.

● One random permutation: (15,12,1,4,2,16,11,8,5,13,6,3,10,7,9,14) corresponds to the genotypes *aA, aA, AA, aA, aa, Aa, Aa, aA,* or $(n_{11}, n_{12}, n_{22}) = (1, 6, 1)$.

PERMUTATION TEST FOR HW

● For 16 elements, there are more than 10^{13} possible permutations. Each random permutation yields a possible sample configuration (n_{11}, n_{12}, n_{22}) whose chi-square can be compared with an observed value, and with a large number of random permuations, the proportion of samples whose chi-square is as large or larger than that observed, approximates the P value.

Example 3 again: The vector to be randomly permuted has 596 + 80 = 676 elements (1,2,3…676) where the integers less than or equal to 596 represent one allele and those greater than 596 represent the other allele. Each successive pair of integers represents a single diploid genotype in the sample.

● Among 1000 random permutations, in 294 cases the chi-square was as large or larger than that observed, yielding $P = 0.294$. The exact value (see above) was 0.290.

● Random permutations are particularly useful when there are more than two alleles and many of them being rare. In such situations rare allele homozygote genotypes may not be in a sample. One practice is to combine all homozygotes and all heterozygotes and compare with the numbers expected under HW.

TWO GENE LOCI

● Two gene loci, *A* and *B*, two alleles at each, A_1 and A_2, with frequencies p_1 and p_2, B_1 and B_2 with frequencies q_1 and q_2

● If mating is random (and other simplifying assumptions of HW, see above) the **genotypes** are expected in **proportions**

A_1A_1 , A_1A_2 ,	p_1^2 , $2p_1p_2$, p_2^2
B_1B_1, B_1B_2 , B_2B_2	q_1^2 , $2q_1q_2$, q_2^2

● It is important to realize that ***within*** each locus the alleles (A_1 , A_2 and B_1 ,B_2) are in random associations, however, the ***alleles at A need not be in randomly associated with alleles at B.***

● Different genes that show randomly associated alleles are said to be in a state of ”linkage equilibrium” and genes not in random association are said to be in ”linkage disequilibrium”. Here ”linkage” has nothing to do with physical linkage of genes.

● With linkage equilibrium the **gametic frequencies** are

$A_1B_1 : p_1q_1$, $A_1B_2 : p_1q_2$, $A_2B_1 : p_2q_1$, $A_2B_2 : p_2q_2$.

● With random mating (and other simplifying assumptions), genes are expected to be in linkage equilibrium.

LINKED TWO LOCI

● Let´s now assume that *A* and *B* and physically linked to each other.

● There are two types of ***double heterozygotes***:

allele combination in one chromosome / the other chromosome

A_1B_1 / A_2B_2

A_1B_2 / A_2B_1

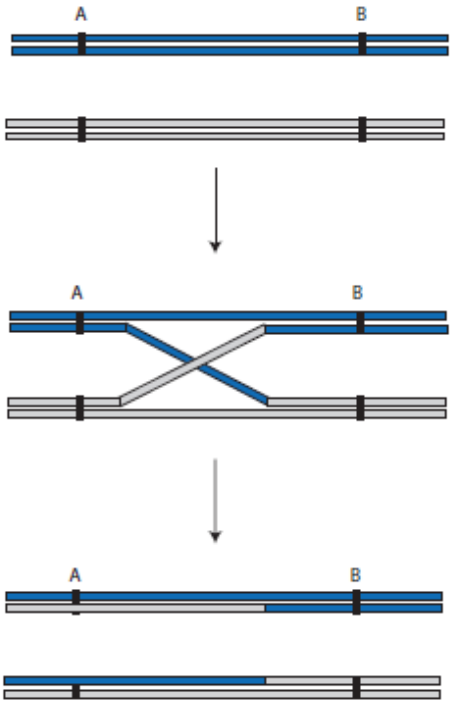
● In the first case the genotype was formed by union of an A_1B_1 gamete with an A_2B_2 gamete and in the second case A_1B_2 gamete with A_2B_1 gamete.

● The genotype A_1B_1 / A_2B_2 produces four different types of gametes to the next generation:

(1) A_1B_1 (2) A_2B_2 **non-recombinant gametes,** alleles associated as they were in the previous generation

(3) A_1B_2 (4) A_2B_1 **recombinant gametes,** alleles are associated differently than they were in the previous generation

A SCHEMATIC PICTURE OF RECOMBINATION BETWEEN GENES A and B



LINKAGE AND "LINKAGE DISEQUILIBRIUM"

- Recall Mendelian segregation: the frequency of gametic type 1 equals that of type 2, and the frequency of gametic type 3 equals that of type 4.

- The overall frequency of recombinant gametes (3 + 4) does not necessarily equal the overall frequency of nonrecombinant gametes (1 + 2).

- **Recombination fraction, *r,***
  refers to the proportion of recombinant gametes produced by a double heterozygote.

- For genes on different chromosomes, $r = 0.5$ because the four possible gametic types are produced in equal frequency.

- For genes on the same chromosome, r depends on their distance apart.

  - In meiosis each chromosome pair exchanges part of chromosomal segments.

  - The closer the two genes are, the less likely is that exchange occurs.

- Suppose, for example, that the genotype *AB/ab* produces gametes *AB, ab, Ab, aB* in proportions 0.38, 0.38, 0.12 and 0.12, respectively. Then the frequency of recombination between the genes is $r = 0.12 + 0.12 = 0.24$

LINKAGE AND "LINKAGE DISEQUILIBRIUM"

● If the frequency of recombination between *A* and *B* genes is *r*, then the genotype *AB / ab* produces the following types of gametes with frequency

AB	(1- *r*) / 2
ab	(1 – *r*) / 2
Ab	*r* / 2
aB	*r* / 2

● Example: Consider two linked genes that have a frequency of recombination of $r = 0.005$. (In the human genome this represents a physical distance of about 5kb.)
What types and frequencies of gametes would be produced by an individual of genotype *AB*/*ab* and an individual of genotype *Ab*/*aB*?
$AB = ab = (1 - 0.005) / 2 = 0.497$, $Ab = aB = 0.005 / 2 = 0.003$
The latter individual produces the same types, but their frequencies are 0.003 and 0.497.

LINKAGE AND "LINKAGE DISEQUILIBRIUM" –> "LINKAGE EQILIBRIUM"

● Consider a **population** in which the frequencies of the chromosome types among the gametes are P_{AB} , P_{Ab} , P_{aB} , P_{ab} and $P_{AB} + P_{Ab} + P_{aB} + P_{ab} = 1$.

● In terms of gamete frequencies the "equilibrium" state is defined as the state in which
$P_{AB} = p_A p_B$, $P_{Ab} = p_A p_b$, $P_{aB} = p_a p_B$, $P_{ab} = p_a p_b$

● Suppose that the genes are not in linkage equilibrium. To determine how rapidly linkage equilibrium is approached: gamete frequencies in the next generation (events with probabilities *r* and 1- *r*) –> frequencies in any generation

$P_{AB}' = (1 - r) P_{AB}$ for the non-recombinants
$+ r\, p_A p_B$ for the recombinants

$P_{AB}' - p_A p_B = (1 - r)(P_{AB} - p_A p_B)$

This equation becomes simplified by defining $D = P_{AB} - p_A p_B$

D_n is the value of D in the nth generation and the equation imples that $D_n = (1 - r) D_{n-1}$

● $D_n = (1 - r) D_{n-1} = (1 - r)^2 D_{n-2} = \ldots\ldots = (1 - r)^n D_0$ where D_0 is the value of D in the founding population ("at the beginning").

● Because $1 - r < 1$, $(1 - r)^n$ goes to zero as n becomes large. How rapidily $(1 - r)^n$ goes to zero depends on r : the closer r is to zero, the slower the rate.

LINKAGE DISEQUILIBRIUM PARAMETER

- The quantity D is called the **linkage disequilibrium parameter** and it can also be written as

$$D = P_{AB} P_{ab} - P_{ab} P_{aB}$$

- Another widely used measure of linkage disequilibrium is related to, but distinct from D.

$$r^2 = D^2 / (p_A q_a p_B q_b)$$

An intuitive biological interpretation of r^2: its square root is the correlation coefficient in allelic state between alleles in the same gamete

STAGE HOME-EXERCISE 1

1. Consider a gene A with alleles A_1 and A_2 at frequencies x_1 and x_2 , and a different gene B in the same population with alleles B_1 , B_2 and B_3 at frequencies y_1 , y_2 and y_3 .
   - What are the expected frequencies of gametes with linkage equilibrium assuming that $x_1 = 0.3$, $y_1 = 0.2$ and $y_2 = 0.3$.

2. For a gene with two alleles, A and a, and another gene in the same population with alleles B and b , let p_A and p_a , p_B and p_b the allele frequencies. Set $p_A = 0.7$ and $p_B = 0.3$.
   - What are the the expected frequencies of all possible gametes assuming linkage equilibrium?
   - What are the expected frequencies of all possible gametes if there is linkage disequilibrium with D equal to 50% of its theoretical maximum?

3. The table below shows the estimated gametic frequencies for the alleles of the genes in five populations. ● For each population, calculate the values of D' and r^2 . ● Which populations show the least amount of linkage disequilibrium? ● Which show the greatest amount of linkage diequilibrium? ● Are there any that show relatively large linkage disequilibrium according to D' but not according to r^2 ?

population	P_{AB}	$P_{A\,b}$	$P_{a\,B}$	$P_{a\,b}$
1	0.2598	0.5362	0.0792	0.1248
2	0.0008	0.0196	0.0694	0.9102
3	0.7332	0.0082	0.0230	0.2356
4	0.2363	0.3029	0.2183	0.2425
5	0.0237	0.3460	0.5574	0.0729

STAGE HOME-EXERCISE 1

4. Suppose that in a population produced by random mating, two loci with two alleles, and frequencies $p_A = p_B = 0.5$, and $D_{AB} = 0.2$. Let half of the individuals be females and half males. The recombination fraction between the loci is 0.3 in females and 0.1 in males.
   - What will D_{AB} be in the offspring generation in terms of D_{AB} in the current generation?
   - What will be the frequency of genotype *AA BB* in the offspring generation?

5. Sampling 100 individuals from a diploid population the following numbers of genotypes at two two-allele loci (*A* and *B*) are observed:

	BB	*Bb*	*bb*
AA	0	25	0
Aa	25	0	25
aa	0	25	0

   - Use a 3 x 3 heterogeneity chi-square to test whether the genotypes at these two loci are distributed independently of each other.
   - See if you can also make an estimate of the linkage disequilibrium D_{AB} between these loci. Is there a discrepancy between these two conclusions? Why, or why not?

6. Suppose a multiple-allele locus with gene frequencies $p_1, p_2, ..., p_n$.
   - In terms of these quantities, after random mating, what fraction of copies of allele A_i occur in heterozygotes?
   - What is the overall fraction of all copies that occur in heterozygotes?

STAGE HOME-EXERCISE 1

7. Given the numbers of the nine genotypes in a sample from a diploid population with two two-allele loci, and assuming that the two loci are unlinked, what are the frequencies of the four gamete types among the haploid gametes produced by this sample?
   - Compute D_{AB} for these gametes in terms of the nine genotype numbers.
   - If the genotypes were sampled from a population produced by random mating, with an unknown true value of D_{AB}, what is the expectation of this estimate of D_{AB} in terms of the true unknown value?
   - If we assume that D_{AB} in the gamete population is estimated by doubling the D_{AB} in the gametic output of our sample, will we be making a biased or an unbiased estimate?

8. Suppose that a chromosome has been duplicated so that where there was once one locus, there are now two unlinked loci, each with two alleles, *A* and *a*. We cannot distinguish which locus contributed *A* or *a* to a given genotype. The two loci are each diploid and they are in linkage equilibrium with each other. At the first locus the gene frequency of *A* is p_1, and at the second locus the gene frequency of *A* is p_2.
   - In terms of those two quantities, what are the expected frequencies of genotypes with 4, 3, 2, 1 and 0 *A*'s?

   Note that we cannot tell the difference between, for example, *AAaa* and *AaAa*, so that they both contribute to the genotypes that have 2 A's.