Case-kontrol studier og genetiske associationsmodeller

www.biostat.ku.dk/~bxc/SDC-courses

Bendix Carstensen

Steno Diabetes Center &
Biostatististisk afdeling, KU

bxc@steno.dk

www.biostat.ku.dk/~bxc

Claus Thorn Ekstrøm

Inst. f. Matematik og Fysik, KVL &
Steno Diabetes Center

ekstrom@dina.kvl.dk

www.matfys.kvl.dk/~ekstrom

December 2002

Logarithms and exponentials

$$10^2 = 10 \times 10$$

$$10^3 = 10 \times 10 \times 10$$

$$10^2 \times 10^3 = 10^5$$

$$10^3/10^2 = 10^1$$

$$(10^3)^2 = 10^6$$

$$10^2/10^2 = 10^0 = 1$$

$$10^2/10^3 = 10^{-1} = 1/10$$

$$10^{1/2} \times 10^{1/2} = 10^1$$

$$10^{1/2} = \sqrt{10}$$

$$10^{0.3010} = 2$$

$$\log_{10}(2) = 0.3010$$

$$10^{0.4771} = 3$$

$$\log_{10}(3) = 0.4771$$

$$10^1 = 10$$

$$\log_{10}(10) = 1$$

Multiplication and division

$$2 \times 3 = 6$$

$$\log_{10}(2) = 0.3010$$
$$\log_{10}(3) = 0.4771$$
$$0.3010 + 0.4771 = 0.7781$$
$$\log_{10}(6) = 0.7781$$

$$10^{0.3010} \times 10^{0.4771} = 10^{0.7781}$$
$$10^{0.7781} = 6$$

In general:

$$\log(xy) = \log(x) + \log(y)$$
$$\log(x/y) = \log(x) - \log(y)$$
$$\log(x^a) = a \log(x)$$
$$\log(1/x) = -\log(x)$$

Natural logarithms: $e = 2.7183$

$$\log_e(e) = 1$$

$$e^{0.6931} = 2$$
$$\log_e(2) = 0.6931$$
$$e^{1.0986} = 3$$
$$\log_e(3) = 1.0986$$

$$2 \times 3 = 6$$
$$e^{0.6931} \times e^{1.0986} = e^{1.7918}$$
$$e^{1.7918} = 6$$

In general:

$$e^x = \exp(x)$$
$$e^x \times e^y = e^{x+y}$$
$$e^x / e^y = e^{x-y}$$
$$(e^x)^y = e^{x \times y}$$
$$1/e^x = e^{-x}$$

Names for the logarithms

Engineers and calculators:

$\log$ is the logarithm to base 10.

$\ln$ is the logarithm to base e, the natural log

Matematicians:

$\log$ is the logarithm to base e, the natural log

$\log_{10}$ is the logarithm to base 10.

Why natural logarithms?

For small values of x (relative to 1):

$$\begin{aligned} e^x &\approx 1+x \\ e^{-x} &\approx 1-x \\ \ln(1+x) &\approx x \\ \ln(1-x) &\approx -x \end{aligned} \quad \Rightarrow \quad \begin{aligned} \ln(1.01) &= 0.01 \\ \ln(0.99) &= -0.01 \\ \ln(1.04) &\approx 0.04 \\ \ln(1.20) &= 0.182 \neq 0.20 \end{aligned}$$

But:

$$\begin{aligned} \log_{10}(1.01) &= 0.4343 \times 0.01 \\ \log_{10}(0.99) &= 0.4343 \times -0.01 \end{aligned}$$

$$\log_{10}(x) = 0.4343 \times \ln(x)$$

Hypothesis tests in statistical analysis

For two populations the hypothesis of equal means is normally formulated as:

$$H_0: \quad \mu_1 = \mu_2 \qquad \Leftrightarrow \qquad \delta = \mu_1 - \mu_2 = 0$$

Statisticians would consider two models:

$$1: \quad \begin{array}{rcl} x_{i1} & \sim & \mathcal{N}(\mu_1, \sigma^2) \\ x_{i2} & \sim & \mathcal{N}(\mu_2, \sigma^2) \end{array} \qquad 2: \quad \begin{array}{rcl} x_{i1} & \sim & \mathcal{N}(\mu, \sigma^2) \\ x_{i2} & \sim & \mathcal{N}(\mu, \sigma^2) \end{array}$$

H_0 would in this context then be:
Can model 1 be reduced to model 2 ?

Hypothesis testing is comparison of models.

Comparing statistical models

- Can a complicated model be reduced to one describing data in a simpler fashion?

  This is the kind of model that one would like to see accepted.

- Can a model be reduced to a model that describes data as not varying with exposure / treatment?

  This is the kind of model that one would like to see rejected.

Relevance of $p < 0.05$ depends on context.

Probability

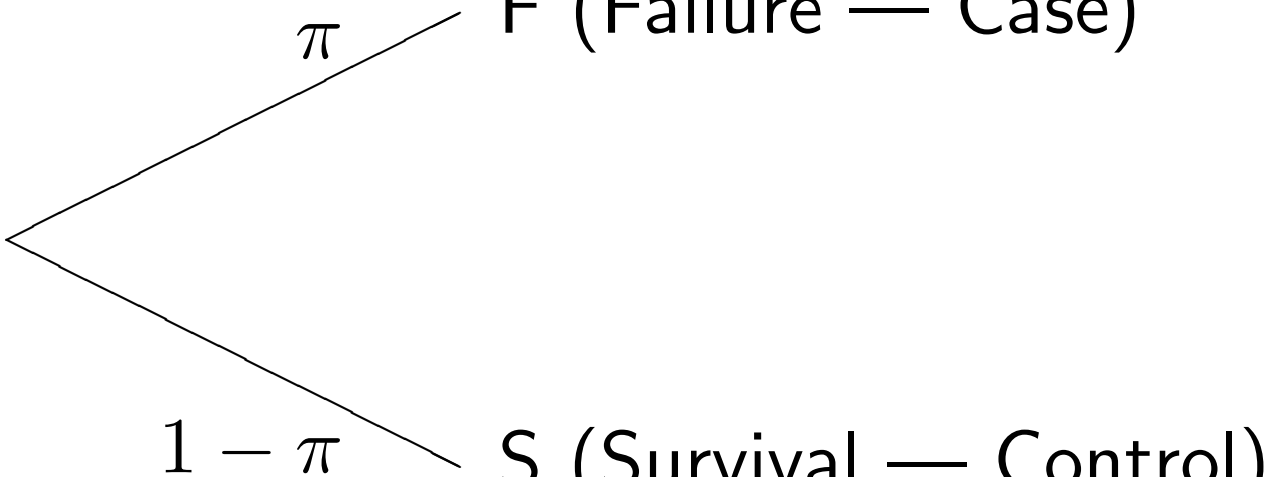
In all scientific studies the outcome is subject to random variation.

In case-control studies and association studies outcomes and exposures are discrete:

- Case / Control
- Genotype: aa / aA / AA

"Measurement"-error described by probabilities for each possible outcome.

The binary probability model

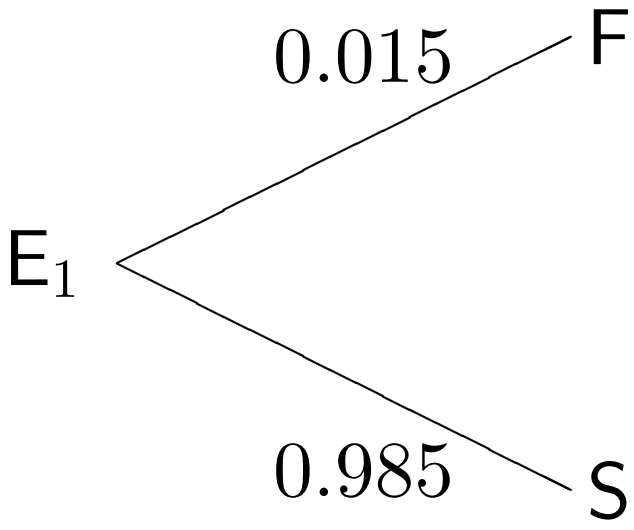


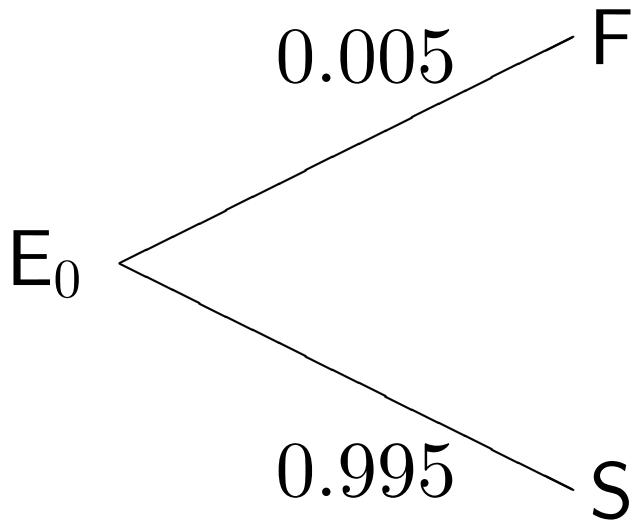
The **risk** parameter: π (pi).

The **odds** parameter: ω (omega).

$$\omega = \frac{\pi}{1-\pi} \quad \Leftrightarrow \quad \pi = \frac{\omega}{1+\omega}$$

Conditional probabilities of failure

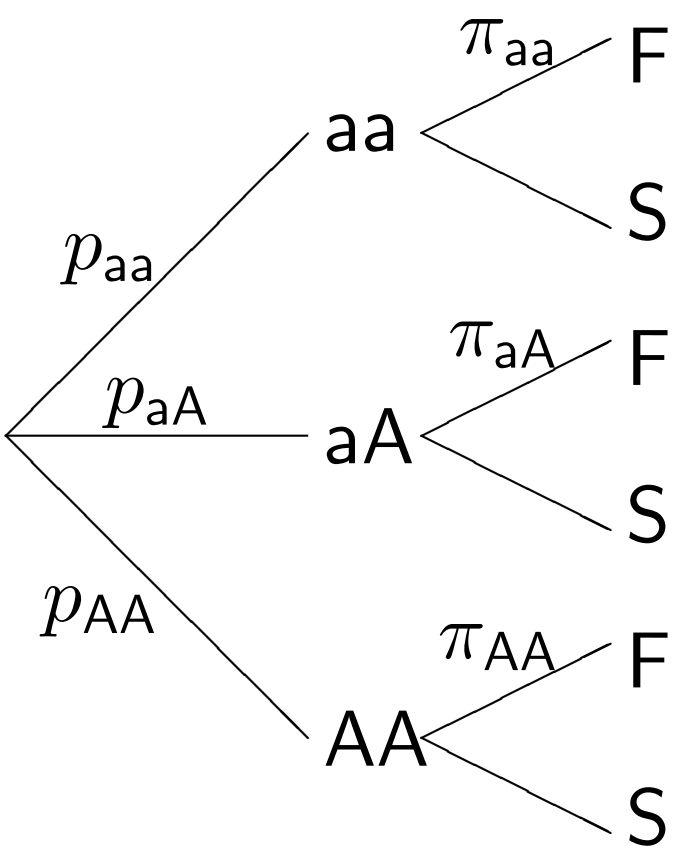




$$\mathrm{P}\{\mathsf{F} \mid \mathsf{E}_1\} = 0.015 \qquad \mathrm{P}\{\mathsf{F} \mid \mathsf{E}_0\} = 0.005$$

Risk for exposed individuals is increased by a factor of $0.015/0.005 = 3.0$, relative to unexposed

Conditional probabilities of failure



p_{aa} is the probability that a person has genotype aa.

π_{aa} is the conditional probability of failure **given** genotype aa.

$p_{\mathsf{aa}} \times \pi_{\mathsf{aa}}$ is the probability that a person has genotype aa **and** fails.

Relationship between follow–up studies and case–control studies

In a **cohort study**, the relationship between exposure and disease incidence is investigated by following the entire cohort and measuring the rate of occurrence of new cases in the different exposure groups.

The follow–up allows the investigator to register those subjects who develop the disease during the study period and to identify those who remain free of the disease.

In a **case-control study** the subjects who develop the disease (the cases) are registered by some other mechanism than follow-up, and a group of healthy subjects (the controls) is used to represent the subjects who do not develop the disease.

Rationale behind case-control studies

- In a follow-up study, rates among exposed and non-exposed are estimated by:

$$\frac{D_1}{Y_1} \qquad \frac{D_0}{Y_0}$$

where D are no. events and Y person-years.

The rate ratio is estimated by:

$$\frac{D_1}{Y_1} \Big/ \frac{D_0}{Y_0} = \frac{D_1}{D_0} \Big/ \frac{Y_1}{Y_0}$$

Necessary to classify both cases and person-years by exposure.

- In a case-control study we use the same cases, but select controls to represent the distribution of risk time between exposed and unexposed:

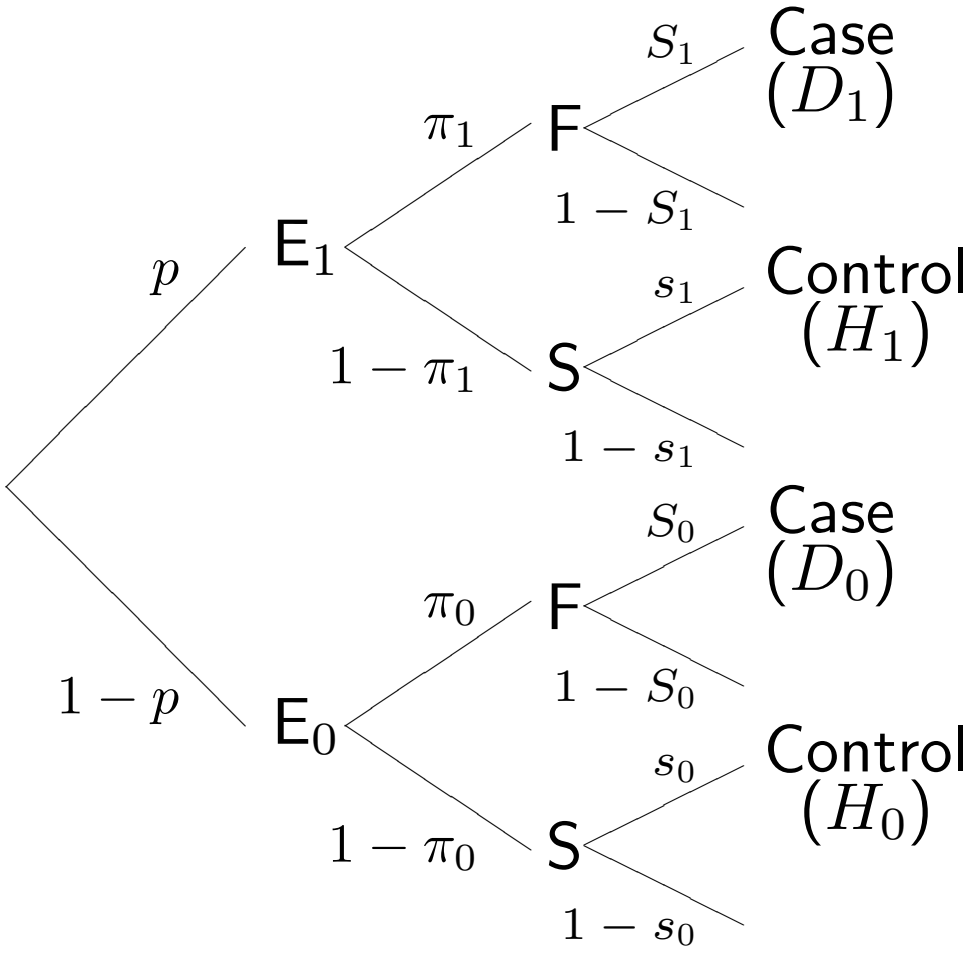
$$\frac{H_1}{H_0} \approx \frac{Y_1}{Y_0}$$

  Therefore the rate ratio is estimated by:

$$\frac{D_1}{D_0} \Big/ \frac{H_1}{H_0}$$

- Controls represent risk time, **not** disease-free persons.

Case–control probability tree



Probability

$p\pi_1 \times S_1$

$p(1-\pi_1) \times s_1$

$(1-p)\pi_0 \times S_0$

$(1-p)(1-\pi_0) \times s_0$

The case-control ratio (disaese odds):

$$\frac{D_1}{H_1} = \frac{S_1}{s_1} \times \frac{\pi_1}{1-\pi_1} \qquad \frac{D_0}{H_0} = \frac{S_0}{s_0} \times \frac{\pi_0}{1-\pi_0}$$

$$\text{Odds-ratio} = \text{OR}_{\text{study}} = \frac{D_1/H_1}{D_0/H_0} = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} = \text{OR}_{\text{population}}$$

but only if $S_1/s_1 = S_0/s_0$, i.e. if sampling fractions are independent of exposure:

$$S_1 = S_0 \qquad \text{and} \qquad s_1 = s_0$$

S sampling fraction for cases — large
s sampling fraction for controls — small

Estimation from case-control study

Odds-ratio of disease between exposed and unexposed *given inclusion in the study*:

$$\mathrm{OR} = \frac{\omega_1}{\omega_0} = \frac{\pi_1}{1-\pi_1} \Big/ \frac{\pi_0}{1-\pi_0}$$

is the same as the odds-ratio of disease between exposed and unexposed *in the "study base"*, **provided** that is the selection mechanism (sampling fractions) is **only** depending on case/control status.

Log-likelihood for case-control studies

Likelihood: Probability of observed data given the statistical model.

Log-Likelihood (conditional on being included) is a binomial likelihood with odds ω_0 and $\omega_1 = \theta\omega_0$

$$D_0 \ln(\omega_0) - N_0 \ln(1 + \omega_0) + D_1 \ln(\theta\omega_0) - N_1 \ln(1 + \theta\omega_0)$$

Odds-ratio (θ) is the ratio of ω_1 to ω_0, so:

$$\ln(\theta) = \ln(\omega_1) - \ln(\omega_0)$$

Estimates of $\ln(\omega_1)$ and $\ln(\omega_0)$ are:

$$\ln\left(\frac{D_1}{H_1}\right) \qquad \text{and} \qquad \ln\left(\frac{D_0}{H_0}\right)$$

with standard errors:

$$\sqrt{\frac{1}{D_1}+\frac{1}{H_1}} \qquad \text{and} \qquad \sqrt{\frac{1}{D_0}+\frac{1}{H_0}}$$

Exposed and unexposed form two independent bodies of data, so the estimate of $\ln(\theta) \quad [= \ln(\mathrm{OR})]$ is

$$\ln\left(\frac{D_1}{H_1}\right) - \ln\left(\frac{D_0}{H_0}\right), \quad \text{s.e.} = \sqrt{\frac{1}{D_1}+\frac{1}{H_1}+\frac{1}{D_0}+\frac{1}{H_0}}$$

Computing c.i. for odds-ratios

$$\hat{\mathrm{OR}} = \frac{D_1/H_1}{D_0/H_0} \qquad \text{s.e.}[\ln(\mathrm{OR})] = \sqrt{\frac{1}{D_1} + \frac{1}{H_1} + \frac{1}{D_0} + \frac{1}{H_0}}$$

95% c.i. for $\ln(\mathrm{OR})$:

$$\ln(\mathrm{OR}) \pm 1.96 \times \text{s.e.}[\ln(\mathrm{OR})]$$

95% c.i. for OR by taking the exponential:

$$\mathrm{OR} \overset{\times}{\div} \underbrace{\exp\left(1.96 \times \text{s.e.}[\ln(\mathrm{OR})]\right)}_{\text{error factor}}$$

Kir 6.2 homozygotes and diabetes

Genotype	Diabetes cases	Population controls
KK	134	124
EE/EK	669	738

What is the odds-ratio of diabetes associated with being homozygous for the K-allele?

This compares KK genotypic persons with EE and EK seen as one group.

How precisely is this odds-ratio determined?

$$\mathrm{OR} = \frac{D_1/H_1}{D_0/H_0} = \frac{134/124}{669/738} = \frac{1.081}{0.907} = 1.192 = 1.19$$

$$\begin{aligned} \mathrm{s.e.}(\ln[\mathrm{OR}]) &= \sqrt{\frac{1}{D_1} + \frac{1}{H_1} + \frac{1}{D_0} + \frac{1}{H_0}} \\ &= \sqrt{\frac{1}{134} + \frac{1}{124} + \frac{1}{669} + \frac{1}{738}} = 0.136 \end{aligned}$$

The 95% limits for the odds-ratio are:

$$\mathrm{OR} \overset{\times}{\div} \exp(1.96 \times 0.136) = 1.192 \overset{\times}{\div} 1.304 = (0.91 - 1.55)$$

K-carriers and diabetes: your turn!

Genotype	Diabetes cases	Population controls
EK/KK	516	532
EE	287	330

What is the odds-ratio of diabetes associated with being a carrier for the K-allele?

This compares KK/EK persons with EE persons.

How precisely is this odds-ratio determined — give a 95% c.i.

Solution to exercise

$$\mathrm{OR} = \frac{D_1/H_1}{D_0/H_0} = \frac{516/532}{287/330} = \frac{0.970}{0.870} = 1.115$$

$$\mathrm{s.e.}(\ln[\mathrm{OR}]) \;=\; \sqrt{\frac{1}{516} + \frac{1}{532} + \frac{1}{287} + \frac{1}{330}} = 0.102$$

The 95% limits for the odds-ratio are:

$$\mathrm{OR} \overset{\times}{\div} \exp(1.96 \times 0.102) = 1.115 \overset{\times}{\div} 1.22 = (0.91 - 1.22)$$

More levels of exposure — genotypes

Genotype	Diabetes cases	Population controls	case/ control odds	OR relative to (0)
EE (0)	287	330	0.870	1.000
EK (1)	382	408	0.936	1.077
KK (2)	134	124	1.081	1.243

The **relationship** of case-control ratios is what matters.

Odds-ratio of diabetes for EK vs. EE is 1.08

Odds-ratio of diabetes for KK vs. EE is 1.24

Odds-ratio of diabetes for KK vs. EK is

Odds-ratio (OR) and rate ratio (RR)

- If the disease probability, π, in the study period is small:

$$\pi = \text{cumulative risk} \approx \text{cumulative rate} = \lambda T$$

with λ the rate and T the study period.

- For small π, $1 - \pi \approx 1$, so:

$$\mathrm{OR} = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} \approx \frac{\pi_1}{\pi_0} \approx \frac{\lambda_1}{\lambda_0} = \mathrm{RR}$$

π small $\Rightarrow$ OR estimate of RR.

Case-control studies and genetic association

Problem: Want to examine if a (biological candidate) gene influences disease status.

Idea: If the gene influences disease status the genotype distribution should be different from cases and controls.

Approach:

- Sample cases and control.
- Genotype all individuals
- Examine whether the genotype distribution is different from cases to control.

Genotype	Wt AA	Het Aa	Hom aa
Diabetics	π_{Wt}	π_{Het}	π_{Hom}
Non-diabetics	π^*_{Wt}	π^*_{Het}	π^*_{Hom}

With $\pi_{Wt} + \pi_{Het} + \pi_{Hom} = 1$ and $\pi^*_{Wt} + \pi^*_{Het} + \pi^*_{Hom} = 1$

Test of homogeneity (i.e., identical genotype distribution) for cases and controls:

$$H_0 : \pi_{Wt} = \pi^*_{Wt},\ \pi_{Het} = \pi^*_{Het}$$

Equivalent:

$$H_0 : \text{OR(Het vs Wt)} = \text{OR(Hom vs. Wt)} = 1$$

Test for homogeneity of genotype distributions (aka *association* or *independence*):

- Likelihood ratio test.
- Chi-square test.
- Fisher's exact test.

Tests asymptotically equivalent.

Rule of thumb: *expected* number of observations ≥ 5 for asymptotics to hold.

$$\frac{\#\text{affection status } \#\text{genotype}}{\#\text{ individuals}}$$

Example:

Genotype	Wt	Het	Hom
Diabetics	10	15	12
Non-diabetics	56	40	10

Results:

OR(Het vs Wt) = 2.10 (0.86 ; 5.15)
OR(Hom vs. Wt) = 6.72 (2.29 ; 19.70)

LR test, $\chi^2(2) = 12.602, p = 0.0018$
$\chi^2(2) = 13.44, p = 0.0012$
Fisher's exact test: $p = 0.0017$

Genetics 101

Recall: The *mode of inheritance* is:

Recessive. If *two* copies of the disease allele are needed before a person becomes affected.

Dominant. If just *one* copy of the disease allele results in a person becoming affected.

Additive. If each copy of the disease allele increases the disease risk (i.e., multiply the OR's by the same amount, ψ)

Note: dominance/recessive are dual terms.

Want to determine the mode of inheritance?

Test for recessive

$$H_{\text{recessive}} : \text{OR(Het vs Wt)} = 1$$

Test for dominance

$$H_{\text{dominance}} : \text{OR(Het vs Hom)} = 1$$

Test for additivity

$$H_{\text{additive}} : \text{OR(Wt vs Het)} = \text{OR(Het vs Hom)}$$

(aka *Co-dominance* or *Multiplicative penetrance model*)

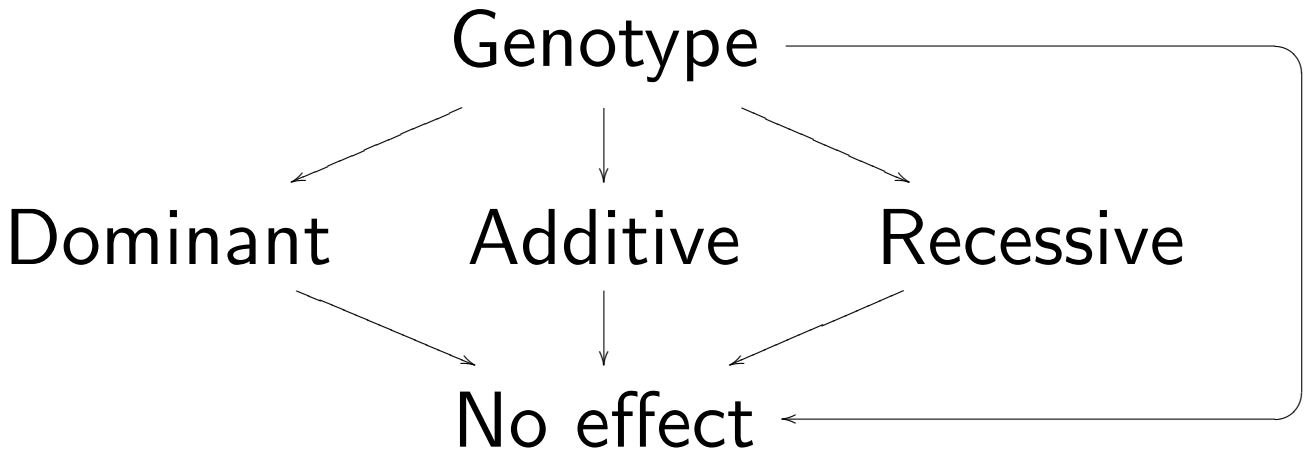
Summary of possible tests

Genotype association

Recessive Additive Dominant

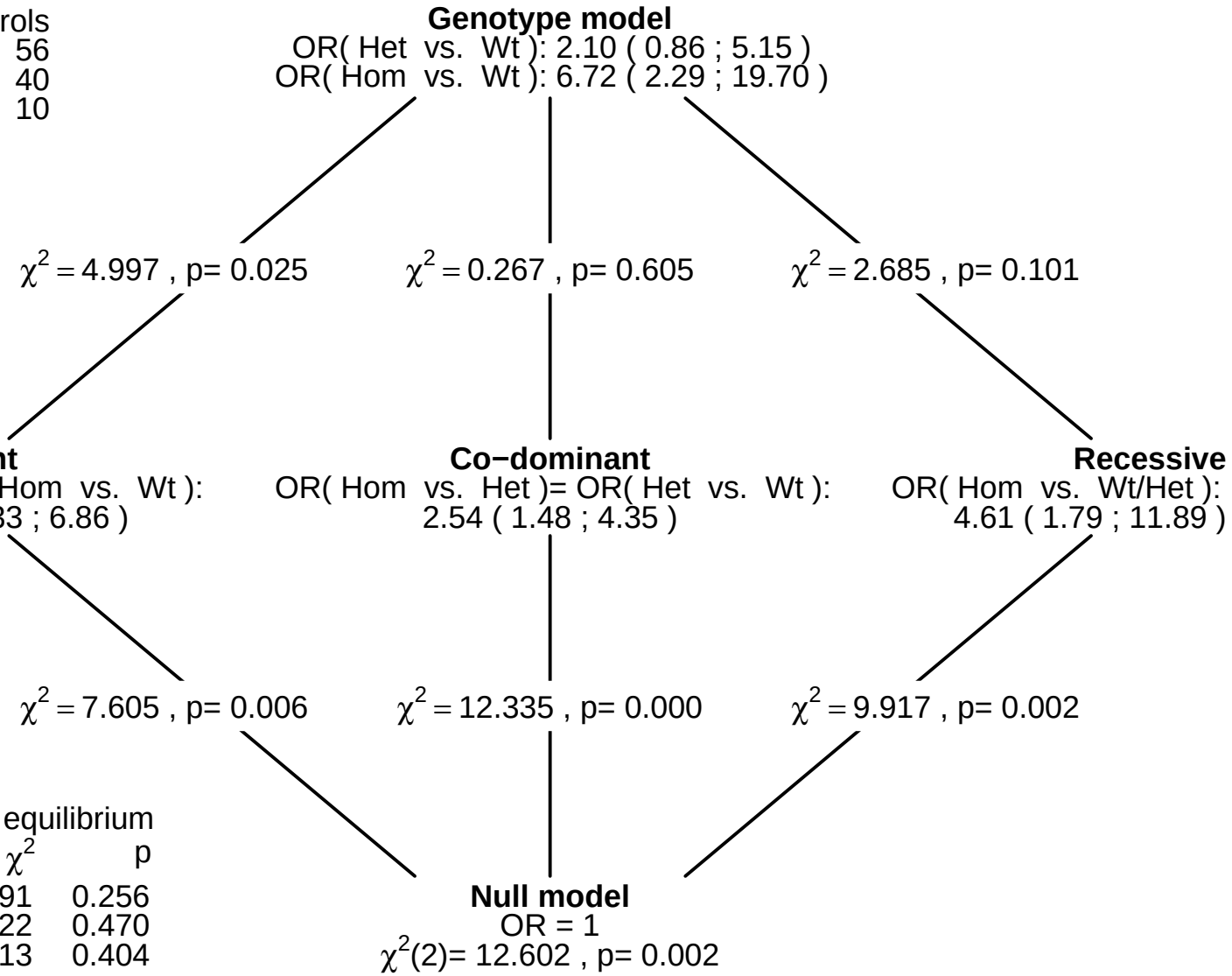
Can still test for "no effect of genotype" after determining mode of inheritance.

Models summarized as follows:



Example

	Cases	Controls
Wt	10	56
Het	15	40
Hom	12	10



Test for H−W equilibrium

	χ^2	p
Cases	1.291	0.256
Controls	0.522	0.470
Both	1.813	0.404

Statistics 101

Test a null hypothesis at significance level $\alpha = 0.05$:

$p < \alpha$ reject the null hypothesis

$p \geq \alpha$ fail to reject the null hypothesis

Bear in mind your null hypothesis when interpreting results!

Generally: focus less on the p-value. The CI of the OR's hold more information!

Common situation:

Test for homogeneity of genotype distributions. $p < 0.05$

We reject the null hypothesis of homogeneity

Test for mode of inheritance.

$p < 0.05$ We reject the null hypothesis of a given mode of inheritance

Hardy-Weinberg equilibrium

A locus is in *Hardy-Weinberg equilibrium* if the frequencies of the genotypes depend *only* on the frequencies of the alleles constituting the genotype (i.e., the two alleles occur *independently* of each other).

Genotype	AA	Aa	aa
General	π_{AA}	π_{Aa}	π_{aa}
HWE	p_A^2	$2p_Ap_a$	p_a^2

Autosomal locus:

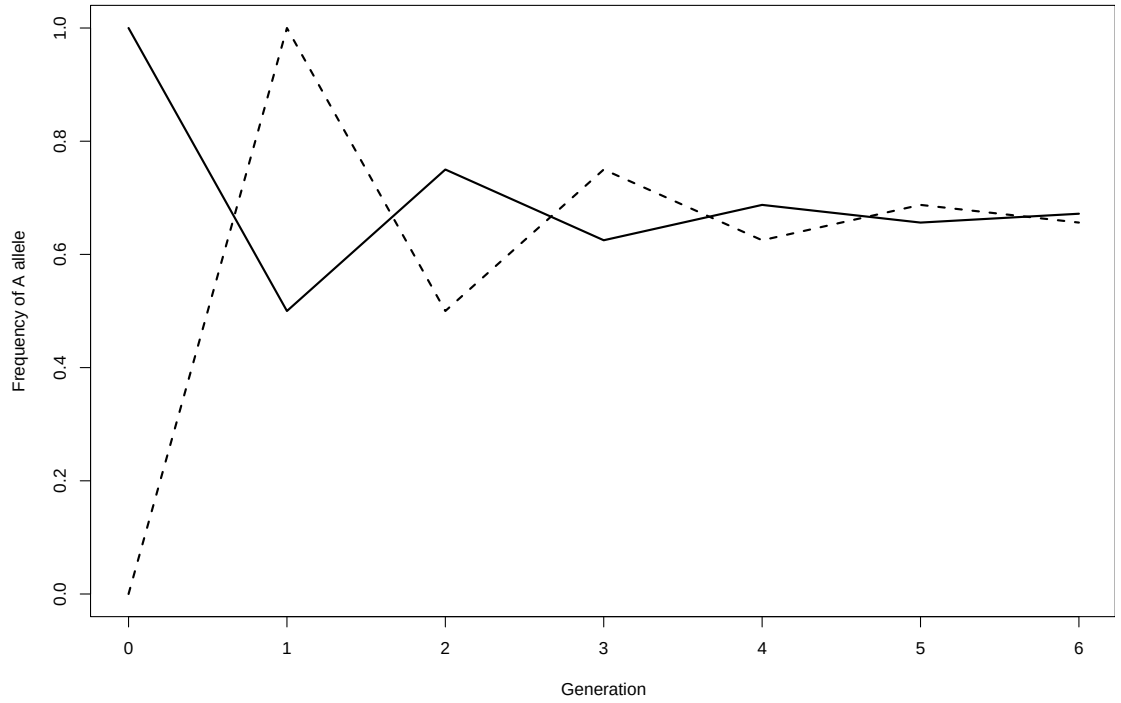
After one generation, an autosomal locus for population exhibiting random mating will be in HWE.

Random mating: each individual selects a mate completely at random.

In practice: mating need only be random with respect to the examined genotype. If the genotype is not related to anything used to choose the mate then random mating satisfied.

X-linked locus:

The distribution of genotypes at an X-linked locus will converge toward HWE.



Deviation from HWE

In general: association between the disease and genotype will result in HWE not being true for cases or controls.

Thus, deviation from HWE preliminary evidence of association.

Exception: additive association does not lead to changes in HWE.

Alternative cause: genotyping errors!

- Ghosting, stuttering (homozygote $\longrightarrow$ heterozygote)
- Allele dropout (heterozygote $\longrightarrow$ homozygote)

Why look at HWE?

The alleles occur independently of genotype — can look at alleles instead of genotypes.

	Wt	Het	Hom
Genotype	AA	Aa	aa
Diabetics	n_{Wt}	n_{Het}	n_{Hom}
Non-diabetics	n^*_{Wt}	n^*_{Het}	n^*_{Hom}

becomes

Allele	A	a
Diabetics	$2 \cdot n_{Wt} + n_{Het}$	$2 \cdot n_{Hom} + n_{Het}$
Non-diabetics	$2 \cdot n^*_{Wt} + n^*_{Het}$	$2 \cdot n^*_{Hom} + n^*_{Het}$

Regular 2×2 table with one OR: OR(A vs a).

Test for no association between allele and disease status:

$$H_0 : \text{OR(A vs a)} = 1$$

To look at alleles (i.e., chromosomes independently) the assumption of HWE is *essential*!

Requirements to consider alleles:

- Cases and controls should be in HWE.
- Test of association is valid if the *population* is in HWE (then under H_0 both cases and controls will be in HWE).

To use standard methods for calculating CI for OR(A vs a) we need

- Rare disease (controls will be in HWE)
- Additive model (cases will also be in HWE)

There are *no* reason to consider alleles instead of genotypes.

PROPOSITION: if both cases and controls are in HWE then the additive model (multiplicative penetrance model) is true. The reverse is not necessarily true.

Complex diseases

Potential problems:

- Broad definition of disease (combination of sub-diseases)
- Random mating (diabetes, obesity)
- Population admixture
- Late onset disease

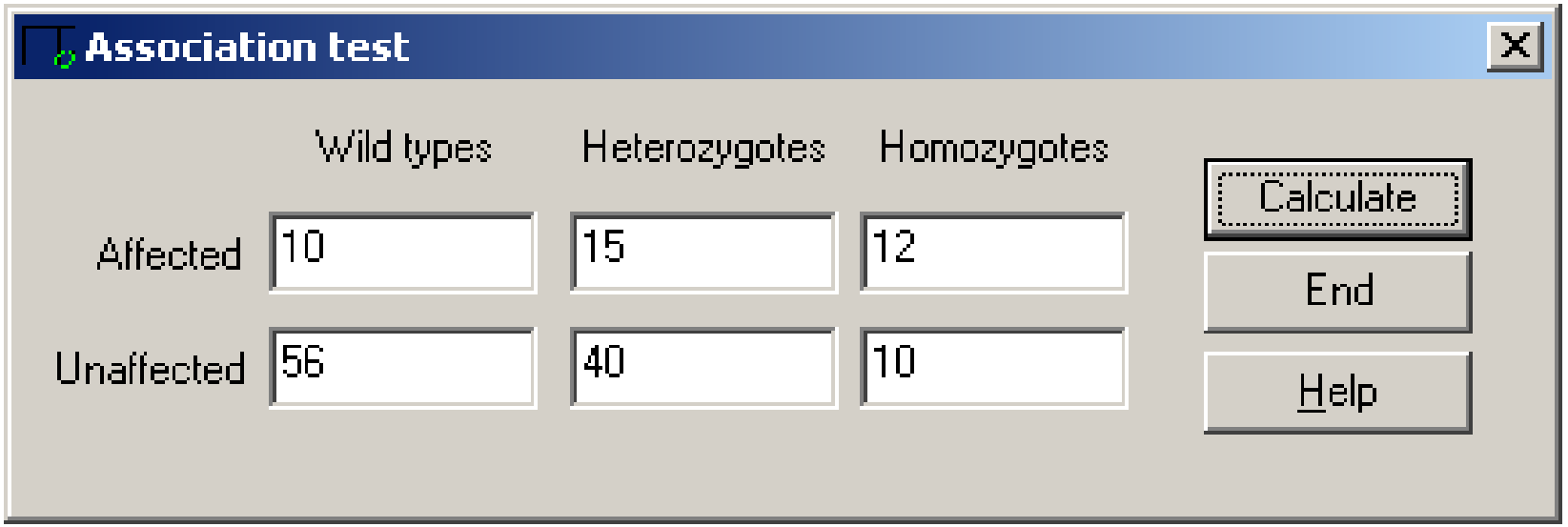
Statistical programs

SPSS, SAS, R Regular statistical problem "easily" solved.

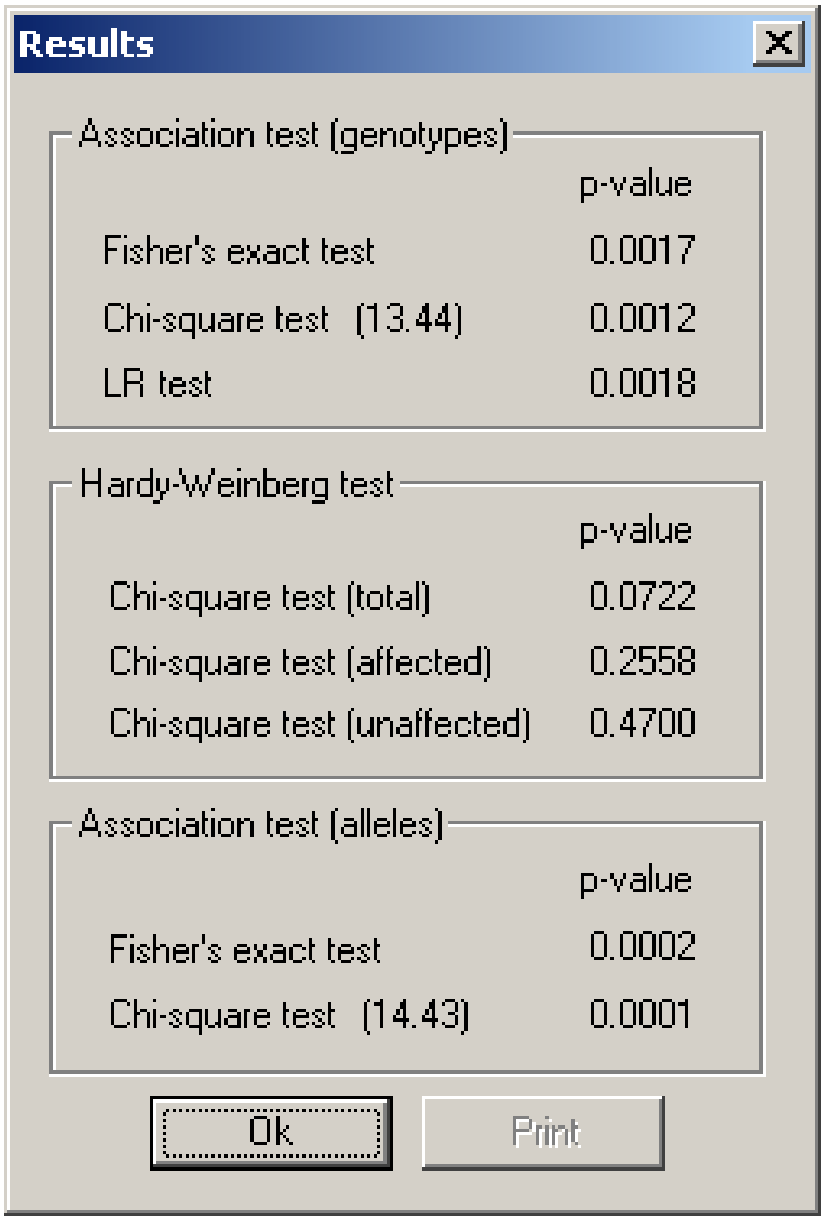
Assotest Windows-based program for genotype/allele association testing.

Web-Assotest Web based program for genotype/allele association testing, HWE, mode of inheritance.

Assotest



Download from `www.ekstroem.com`


Results
Association test (genotypes)
p-value
Fisher's exact test 0.0017
Chi-square test (13.44) 0.0012
LR test 0.0018
Hardy-Weinberg test
p-value
Chi-square test (total) 0.0722
Chi-square test (affected) 0.2558
Chi-square test (unaffected) 0.4700
Association test (alleles)
p-value
Fisher's exact test 0.0002
Chi-square test (14.43) 0.0001
Ok
Print

Fisher's exact test

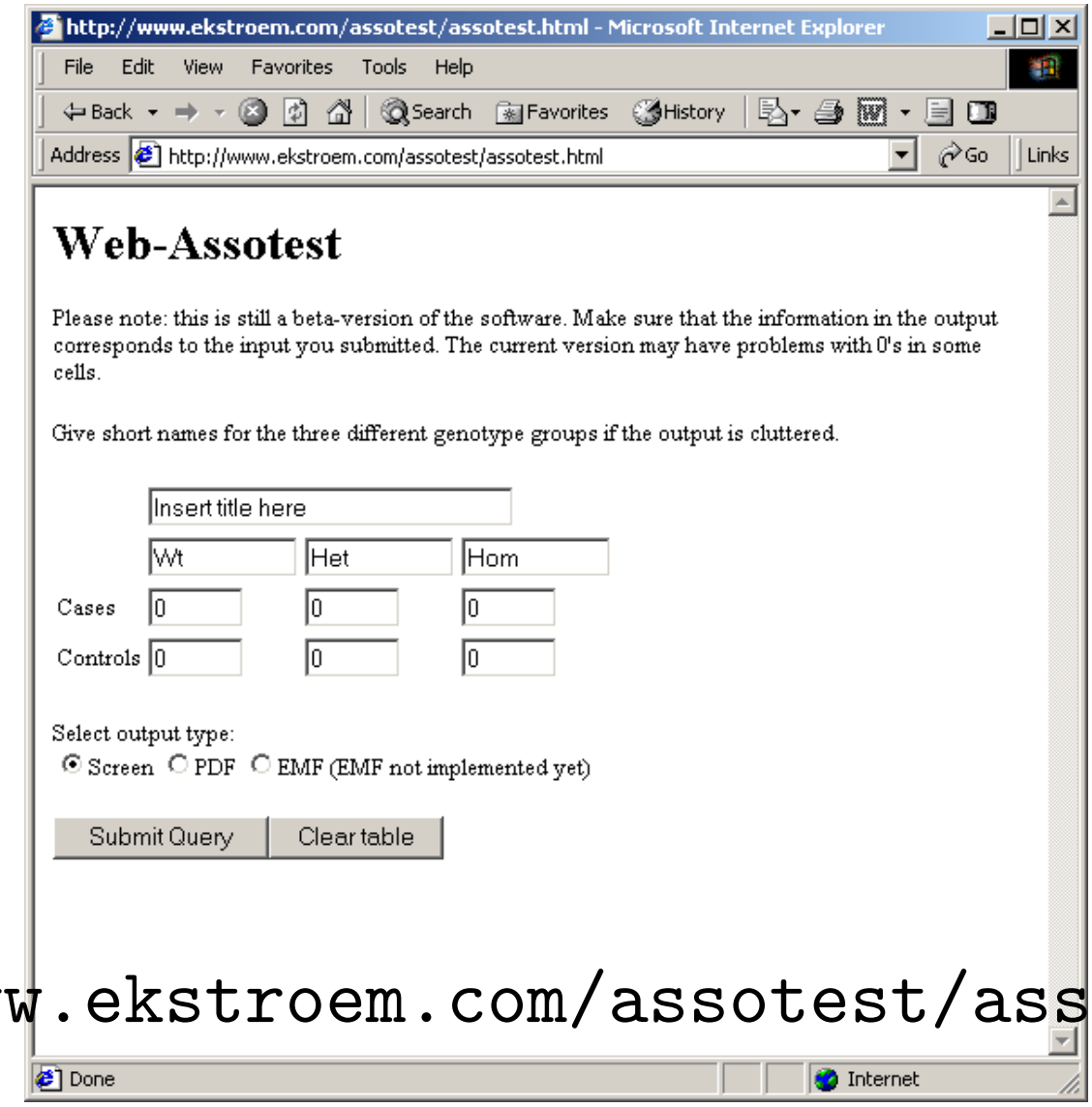
Advantages:

- Exact — don't worry about asymptotics

Disadvantages:

- Provides no information about the relationship of genotype effects
- Computationally intensive

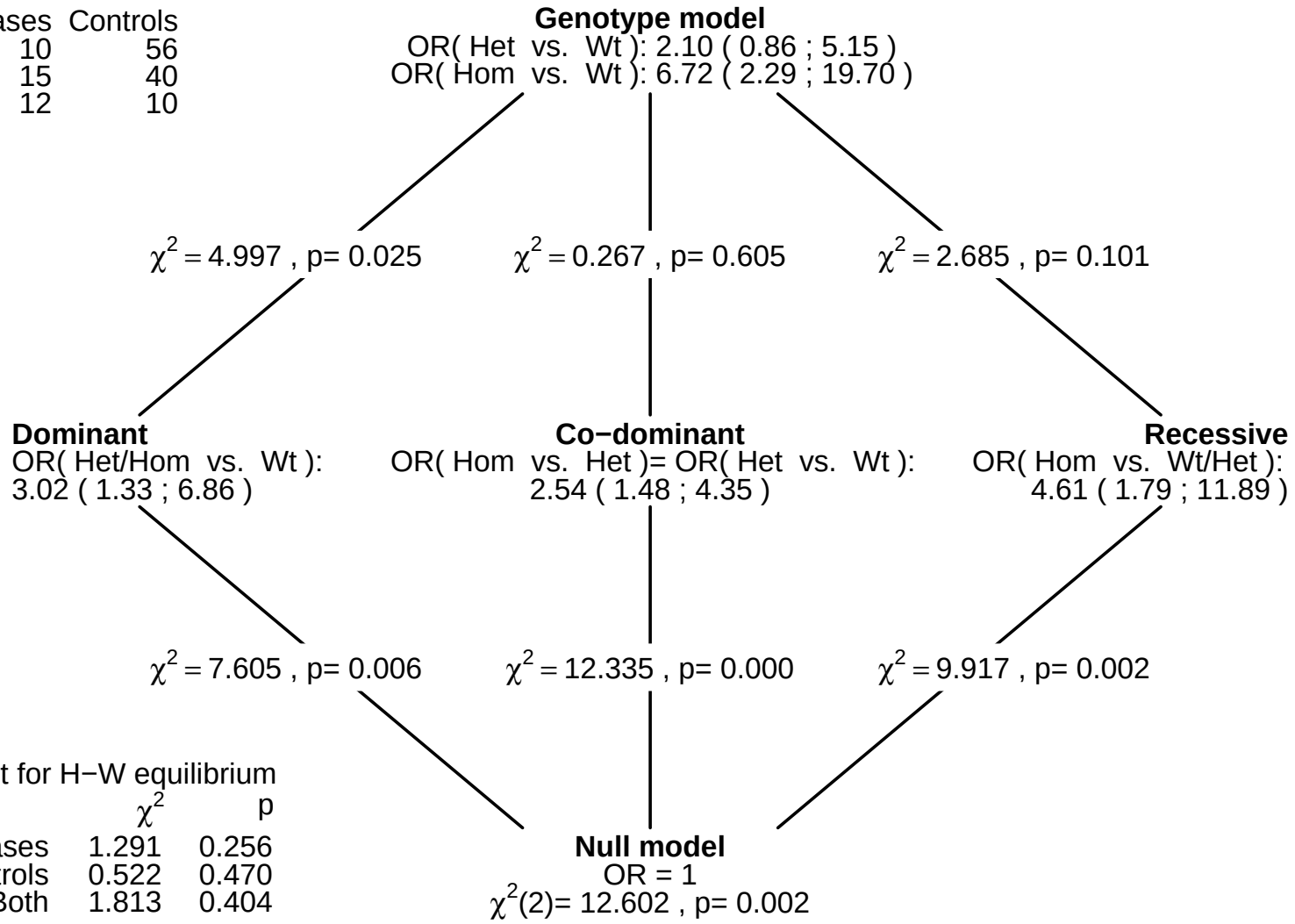
Web-Assotest



Address: `www.ekstroem.com/assotest/assotest.html`

Example

	Cases	Controls
Wt	10	56
Het	15	40
Hom	12	10



Test for H−W equilibrium

	χ^2	p
Cases	1.291	0.256
Controls	0.522	0.470
Both	1.813	0.404

Direct comparison

Assotest	**Web-assotest**
Windows	Web based
Genotype/allele association	All models
Exact/asymptotic tests	Asymptotic tests
p-values only	p-values and CI
HWE	HWE
Total population HWE	Simultaneous HWE

Opgave 1

Vis følgende sætning:

Hvis både cases og kontroller er i HWE er kravene til den additive model opfyldt (altså at OR(Wt vs Het) = OR(Het vs Hom)).

Giv et eksempel, der viser, at det modsatte ikke behøver være sandt.

Opgave 2

Betragt nedenstående datasæt

Genotype	AA	Aa	aa
Diabetics	10	15	12
Non-diabetics	56	40	10

Udregn allelfrekvenser og konfidensintervaller for p_A og p_a.

Confounding

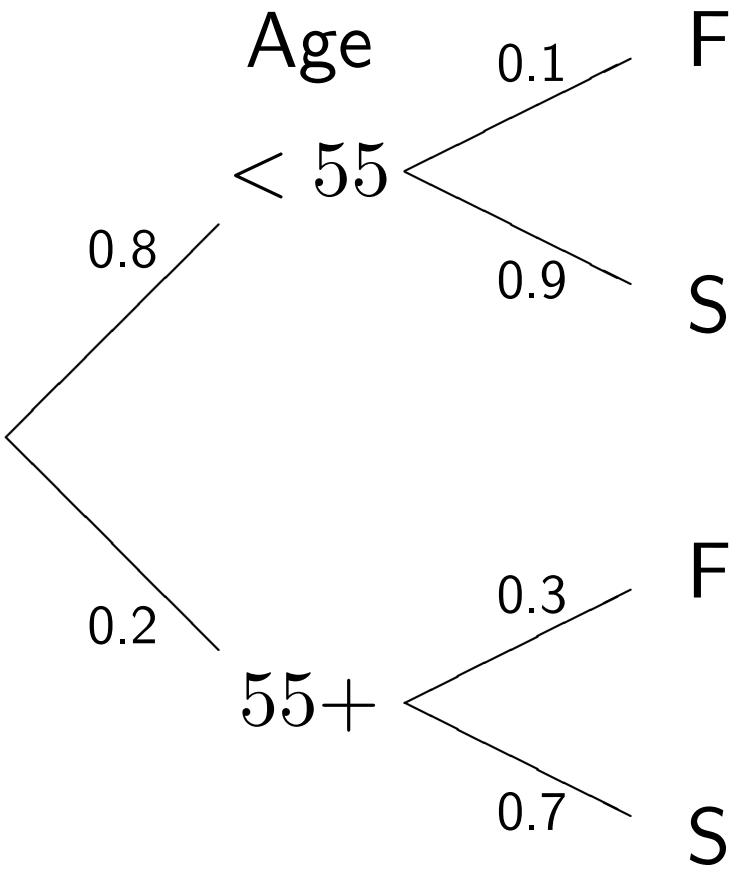
- Epidemiology relies on *observational studies* of *experiments of nature*

- Often these are poor experiments
  — no control for *confounding* by extraneous influences

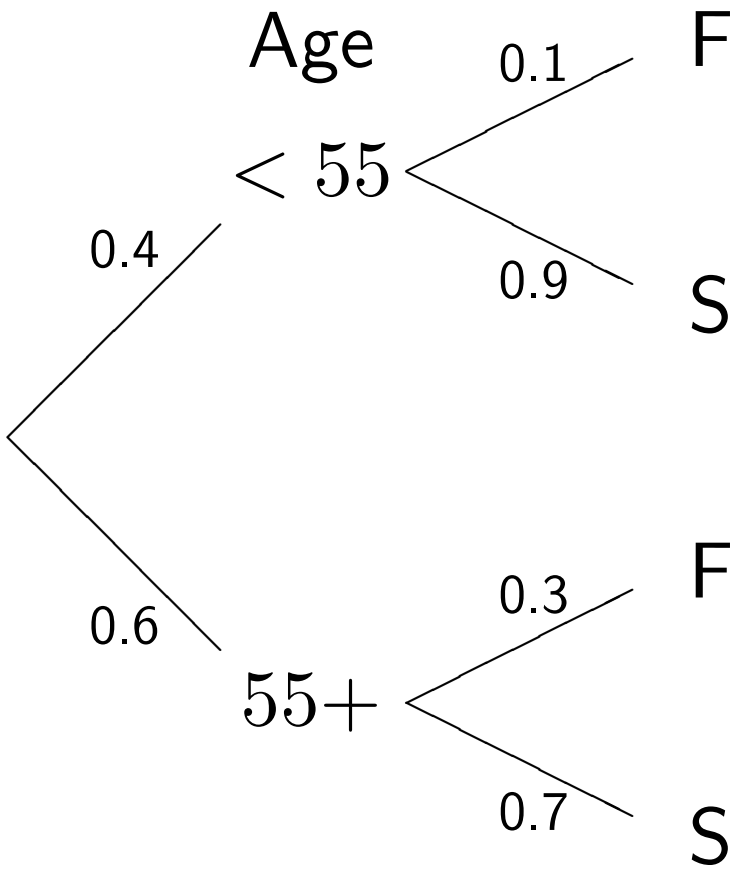
- Definition:

  A confounder is a variable whose influence we would have controlled if we had been able to design the natural experiment.

Example: confounding by age



Unexposed subjects



Exposed subjects

- Probability of failure for **unexposed**:

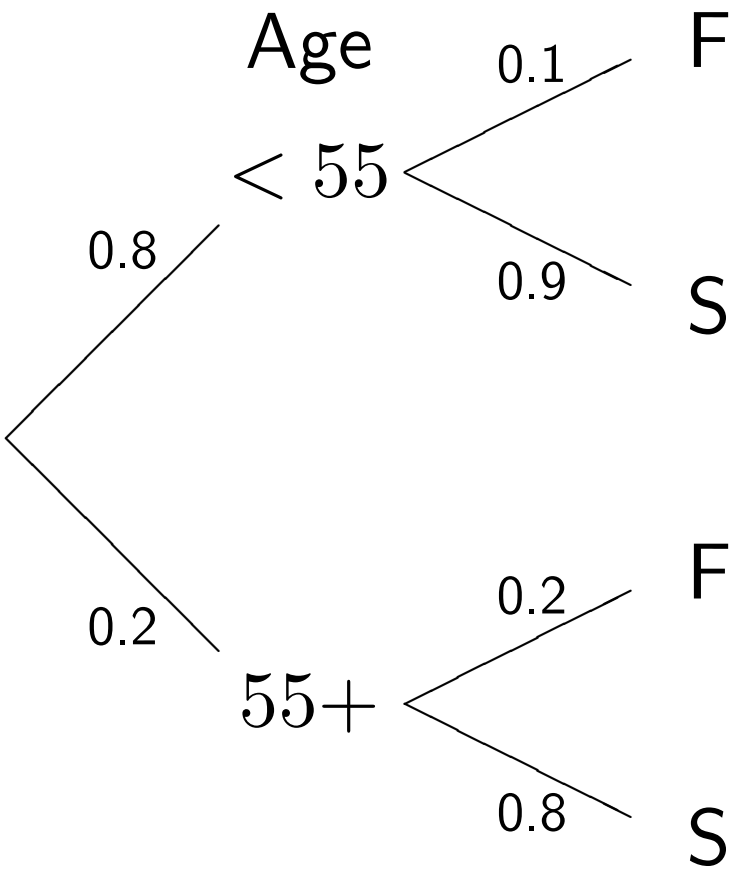
$$(0.8 \times 0.1) + (0.2 \times 0.3) = 0.14$$

- Probability of failure for **exposed**:

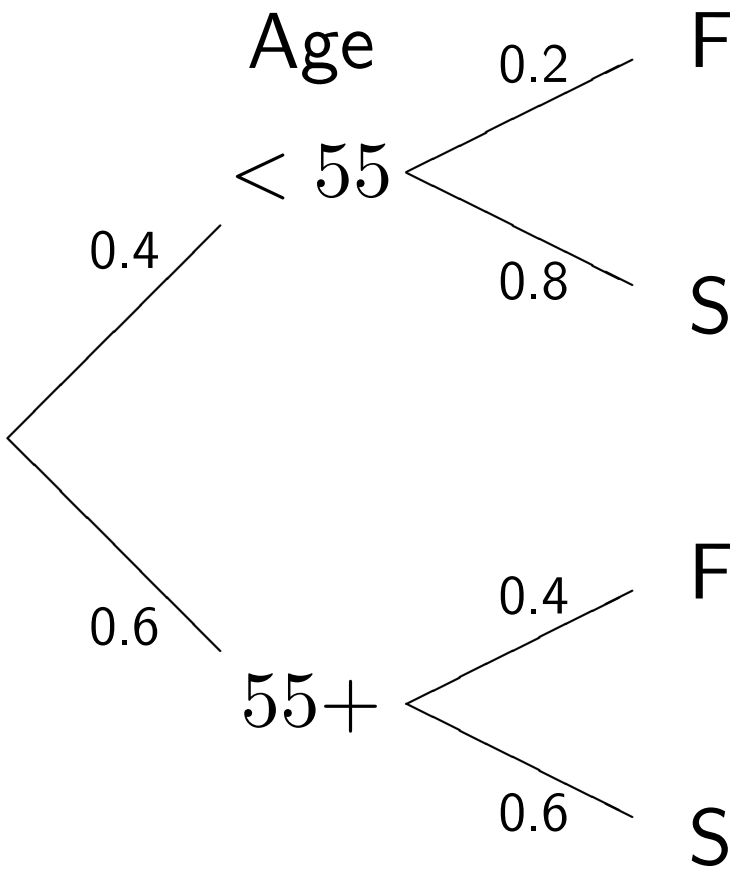
$$(0.4 \times 0.1) + (0.6 \times 0.3) = 0.22$$

- Difference entirely due to difference in age structure.
- When there is a true effect, its magnitude can be distorted by such influences.

Confounding when $\mathrm{RR} = 2$



Unexposed subjects



Exposed subjects

- The true relative risk, $\mathrm{RR}_T = 0.2/0.1 = 0.4/0.2 = 2$
- Probability of failure for **unexposed**:

$$(\quad \times \quad) + (\quad \times \quad) =$$

- Probability of failure for **exposed**:

$$(\quad \times \quad) + (\quad \times \quad) =$$

- The apparent relative risk:

$$\mathrm{RR}_O =$$

- The true relative risk, $\mathrm{RR}_T = 0.2/0.1 = 0.4/0.2 = 2$
- Probability of failure for **unexposed**:

$$(0.8 \times 0.1) + (0.2 \times 0.2) = 0.12$$

- Probability of failure for **exposed**:

$$(0.4 \times 0.2) + (0.6 \times 0.4) = 0.32$$

- The apparent relative risk:

$$\mathrm{RR}_O = 0.32/0.12 = 2.67$$

Confounding

A confounder is:

- Associated with outcome:
  The older persons have higher disease probability.
- Associated with the exposure:
  The older persons are more / less likely to be exposed.
- Is not a result of either exposure or disease.
  Not a statistical property. Cannot be seen from tables.
- Common sense is required!

Controlling confounding

In **controlled experiments** there are two ways of controlling confounding:

1. **Randomization** of subjects to experimental groups so that the *distributions* of the confounder are the same.

2. Hold the confounder **constant**.

Standardization is a statistical technique for controlling for extraneous variables in the analysis of an observational study:

- **Direct** standardization simulates randomization by equalising the distribution of extraneous variables.
- **Indirect** standardization simulates the second method: holding extraneous variables constant.

The latter is the preferred technique in observational studies. It leads to proper statistical modelling.

Indirect standardization

- Aim is to hold age (the confounder) constant.
- Compare exposed and unexposed $within\ age\ strata$
- But this leads to $several$ experiments, each one rather small, hence imprecise.
- Calculate a single $combined$ estimate of the exposure effect over all strata.
- This procedure implies a **model** in which there is no (systematic) variation of effect over strata.

Confounding by age in genetic studies

- Age is associated with outcome — disease, in this case diabetes.

- Age is associated with exposure — genotype, **only** if genotype is associted with mortality.

  Otherwise the genotype distribution will be similar in all age-groups.

Age is not likely to be a confounder in genetic association studies.

Meta-analysis

If several case-control studies are conducted in different populations, they cannot be regarded as one because:

- Study population may be associated with outcome — in this case occurrence of diabetes.
- Study population may be associated with exposure — in this case genotype distribution.

Thus study population should be regarded as a confounder.

Model for confounder control

Assumption of similar effect across studies in different populations: $\mathrm{OR}_p = \theta$ independent of p, so for odds of disease ω_{p1}:

$$\omega_{p1} = \theta\omega_{p0}$$

Odds of disease increase by the same amount, θ, by exposure, regardless of study.

But the disease odds among unexposed, ω_{p0}, may vary between studies.

On the log-scale:

$$\ln\left(\frac{\pi_{p1}}{1-\pi_{p1}}\right) = \ln(\omega_{p1}) = \ln(\theta) + \ln(\omega_{p0})$$

Model for case-control studies

Case-control studies has different sampling fractions for cases (S, large) and controls (s, small):

$$\begin{aligned}
\ln[\text{odds(case} \mid \text{incl.},p)] &= \ln\left[\frac{\pi_{p1}}{1-\pi_{p1}} \times \frac{S_p}{s_p}\right] \\
&= \ln\left[\frac{\pi_{p1}}{1-\pi_{p1}}\right] + \ln\left[\frac{S_p}{s_p}\right] \\
&= \ln(\theta) + \underbrace{\ln(\omega_{p1}) + \ln\left[\frac{S_p}{s_p}\right]}_{\text{intercept, population}}
\end{aligned}$$

Logistic regression model with effects of exposure and study population. Estimates for effect of population is irrelevant, since sampling fractions most likely depends on population.

But population **must** be in the model.

The model with

- exposure (genotype)
- confounder (study population)

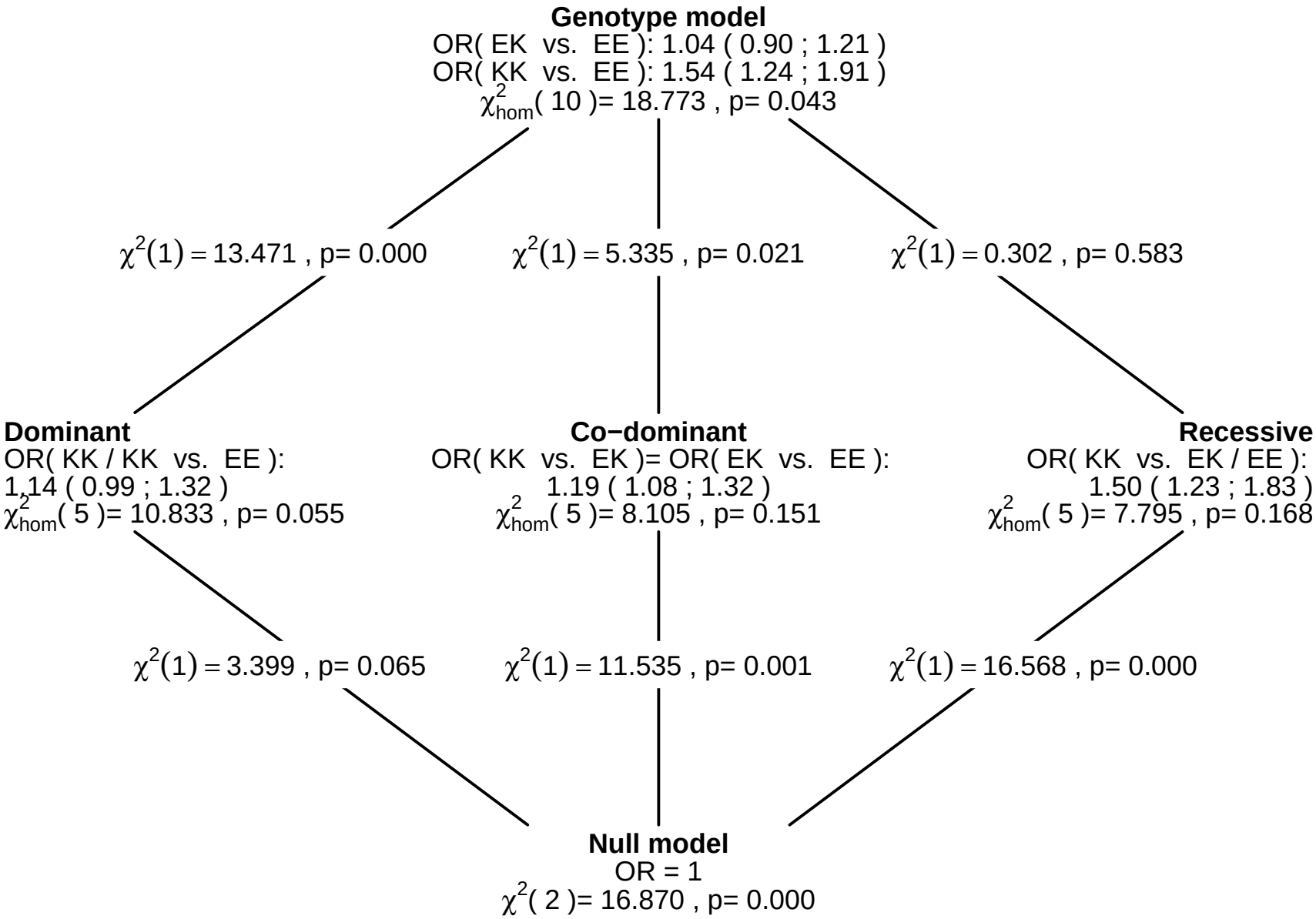
is the meta-analysis model.

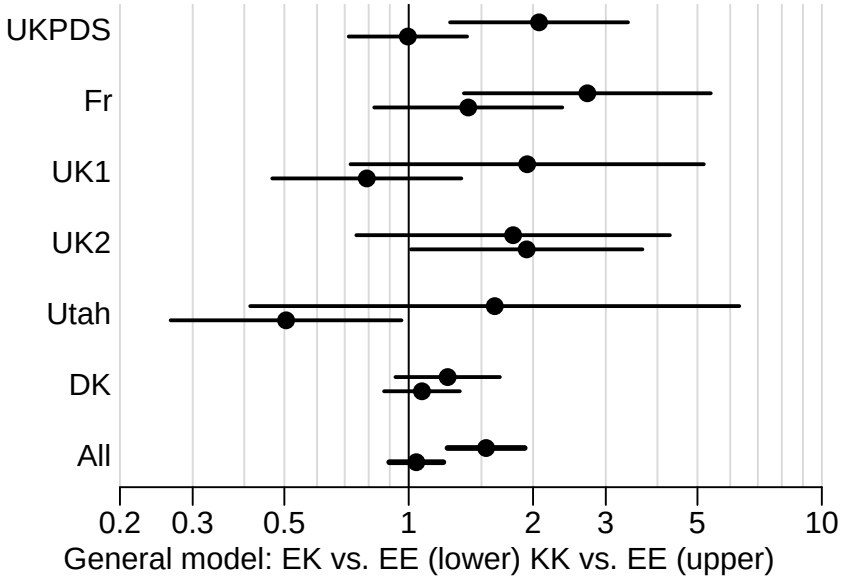
Meta-analysis

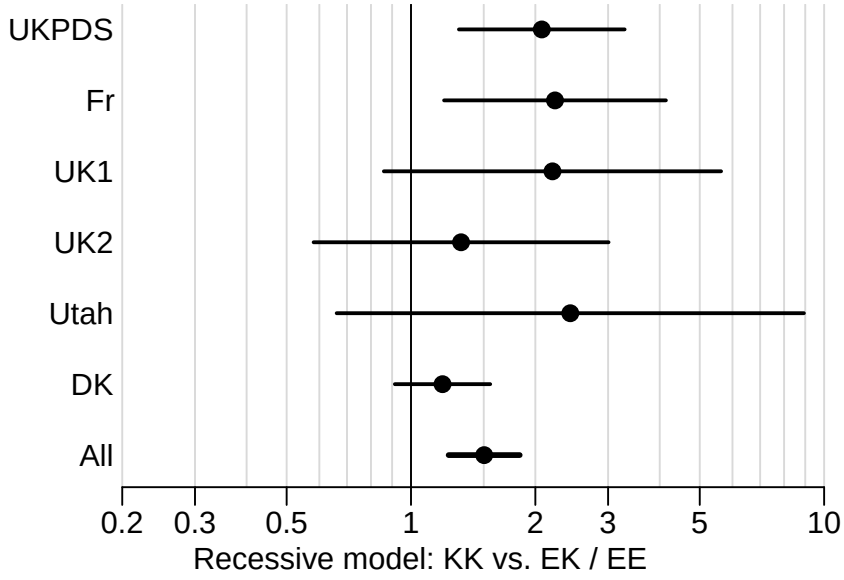
Analysis with study population as controlling variable.
Still two things to consider:

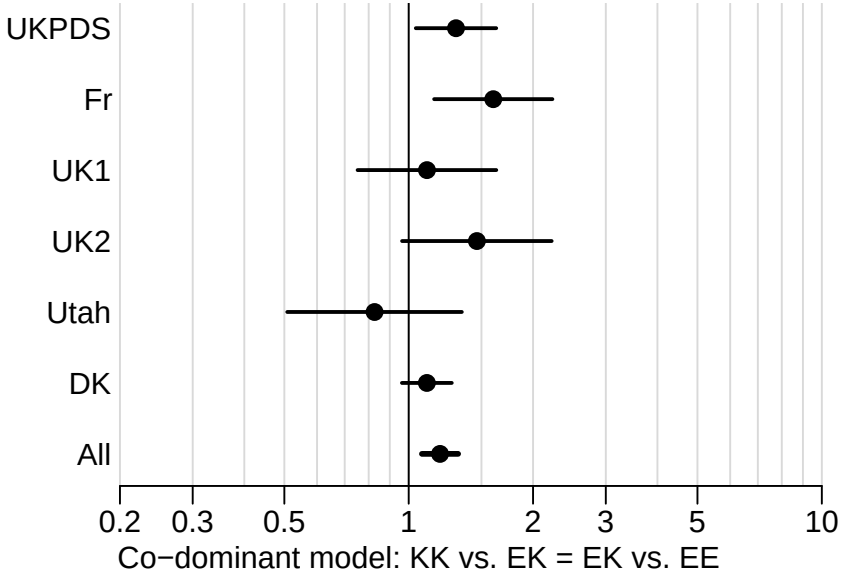
- How is the genotype effect: dominant, co-dominant or recessive?
  Similar to the analysis for one population. But in stratified model.

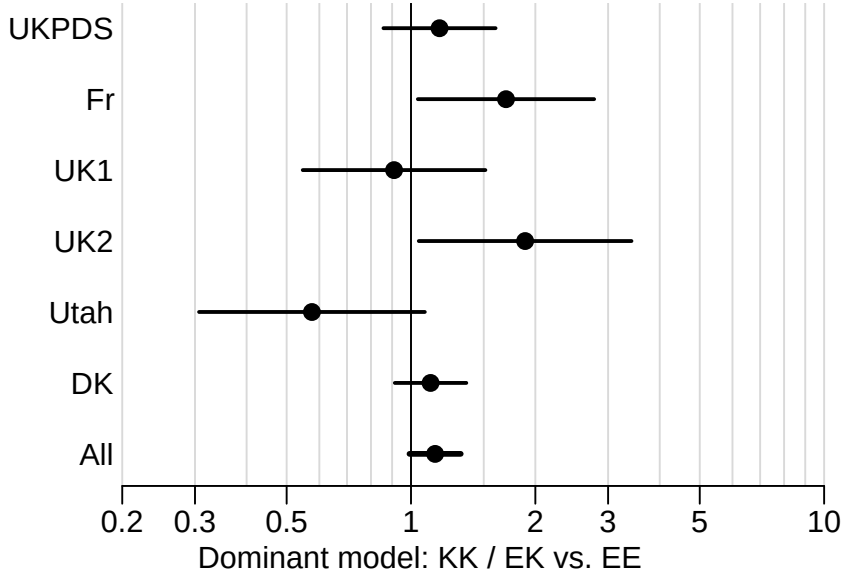
- Is the effect same across populations?
  Test for homogeneity of effect (interaction)


Genotype model
OR(EK vs. EE): 1.04 (0.90 ; 1.21)
OR(KK vs. EE): 1.54 (1.24 ; 1.91)
χ^2_{hom}(10)= 18.773 , p= 0.043
$\chi^2(1) = 13.471$, p= 0.000
$\chi^2(1) = 5.335$, p= 0.021
$\chi^2(1) = 0.302$, p= 0.583
Dominant
OR(KK / KK vs. EE):
1.14 (0.99 ; 1.32)
χ^2_{hom}(5)= 10.833 , p= 0.055
Co−dominant
OR(KK vs. EK)= OR(EK vs. EE):
1.19 (1.08 ; 1.32)
χ^2_{hom}(5)= 8.105 , p= 0.151
Recessive
OR(KK vs. EK / EE):
1.50 (1.23 ; 1.83)
χ^2_{hom}(5)= 7.795 , p= 0.168
$\chi^2(1) = 3.399$, p= 0.065
$\chi^2(1) = 11.535$, p= 0.001
$\chi^2(1) = 16.568$, p= 0.000
Null model
OR = 1
χ^2(2)= 16.870 , p= 0.000


UKPDS
Fr
UK1
UK2
Utah
DK
All
0.2
0.3
0.5
1
2
3
5
10
General model: EK vs. EE (lower) KK vs. EE (upper)


UKPDS
Fr
UK1
UK2
Utah
DK
All
0.2
0.3
0.5
1
2
3
5
10
Recessive model: KK vs. EK / EE


UKPDS
Fr
UK1
UK2
Utah
DK
All
0.2
0.3
0.5
1
2
3
5
10
Co−dominant model: KK vs. EK = EK vs. EE


UKPDS
Fr
UK1
UK2
Utah
DK
All
0.2
0.3
0.5
1
2
3
5
10
Dominant model: KK / EK vs. EE

Do the studies actually show the same?

Apart from the visual inspection of the diagram, formal tests for the models separately may be of interest.

Look at the top left corner in the 2×2 figure layout.

Which studies support a dominant / co-dominant / recessive model?

Is this consonant with what you see in the next tables?

Test for models, single studies

χ^2	Model			
	Dominant	Co-dominant	Recessive	d.f.
UKPDS	9.133	4.759	0.001	1
Fr	4.116	0.451	1.543	1
UK1	3.655	3.521	0.758	1
UK2	0.027	1.234	4.051	1
Utah	3.480	5.923	4.456	1
DK	0.999	0.115	0.470	1
All	21.411	16.003	11.280	6

Test for models, single studies

p-values	Model		
	Dominant	Co-dominant	Recessive
UKPDS	0.003	0.029	0.979
Fr	0.042	0.502	0.214
UK1	0.056	0.061	0.384
UK2	0.869	0.267	0.044
Utah	0.062	0.015	0.035
DK	0.317	0.735	0.493
All	0.002	0.014	0.080

Two different kinds of tests for Dominant / Co-dominant / Recessive:

- Test in stratified model, assuming **same** effect in all populations.

  This is the test shown in the diagram.

- Test in separate models added up.

  Tests for mode of action, allowing for separate effects between populations.

  This is the test in the last line of the table. (Not default output of the meta-analysis program).