# Data Science in Bioinformatics

Palle Villesen & Thomas Bataillon

Does it (the **model**) **fit** with the **data**?
Random .. how?
Week 05 in DS in Bioinformatics

REMEMBER LAST WEEK ?

Hypothesis testing

The binomial distribution

X records number of ” A” in n independent trials that each have 2 outcomes (A/B)

P(X) = ….

**$E(X) = n\ p_A$**

**$V(X) = n\ p_A\ (1-p_A)$ (useful!!!!)**

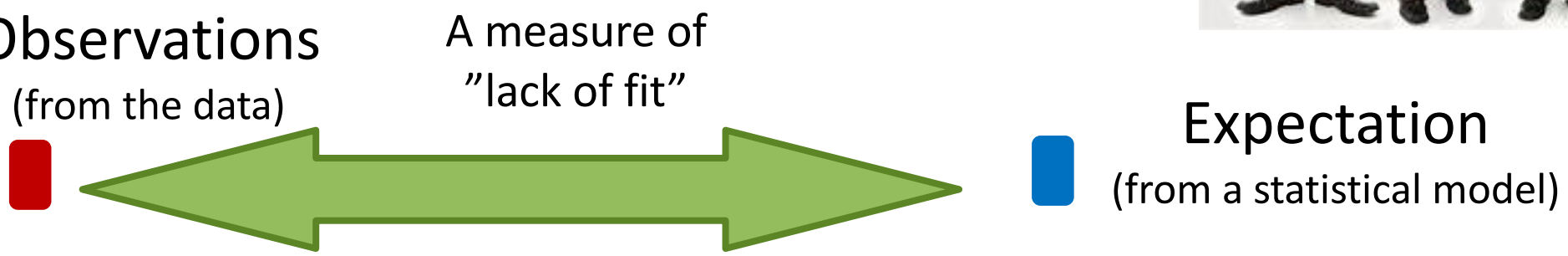
TODAY

Goodness of fit (GOF) tests

- Why it matters ?
- How do they work ?
- What is the ”chi square” **thing anyway ?**

**Random things what does it mean?**

# Goodness-of-fit (GOF) tests: the big picture




A cornerstone of inferential statistics is confronting a (statistical) model with empirical data

In a nutshell:

1. **We measure how well/badly** observed data **fits** the expectations under a null model (H0)

…. Yes that is like a test statistic …

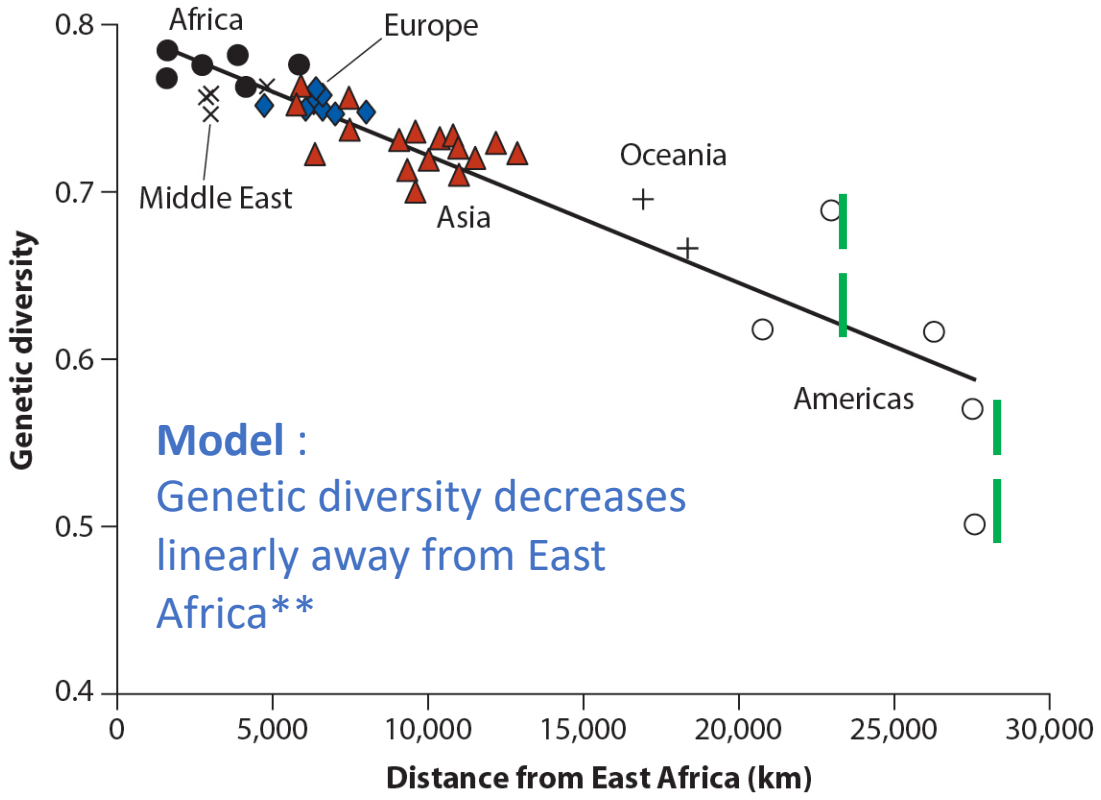
2. **We calculate/simulate the lack of fit expected just by chance (H0 distribution)**

3. **We take a (statistical) decision** ☺

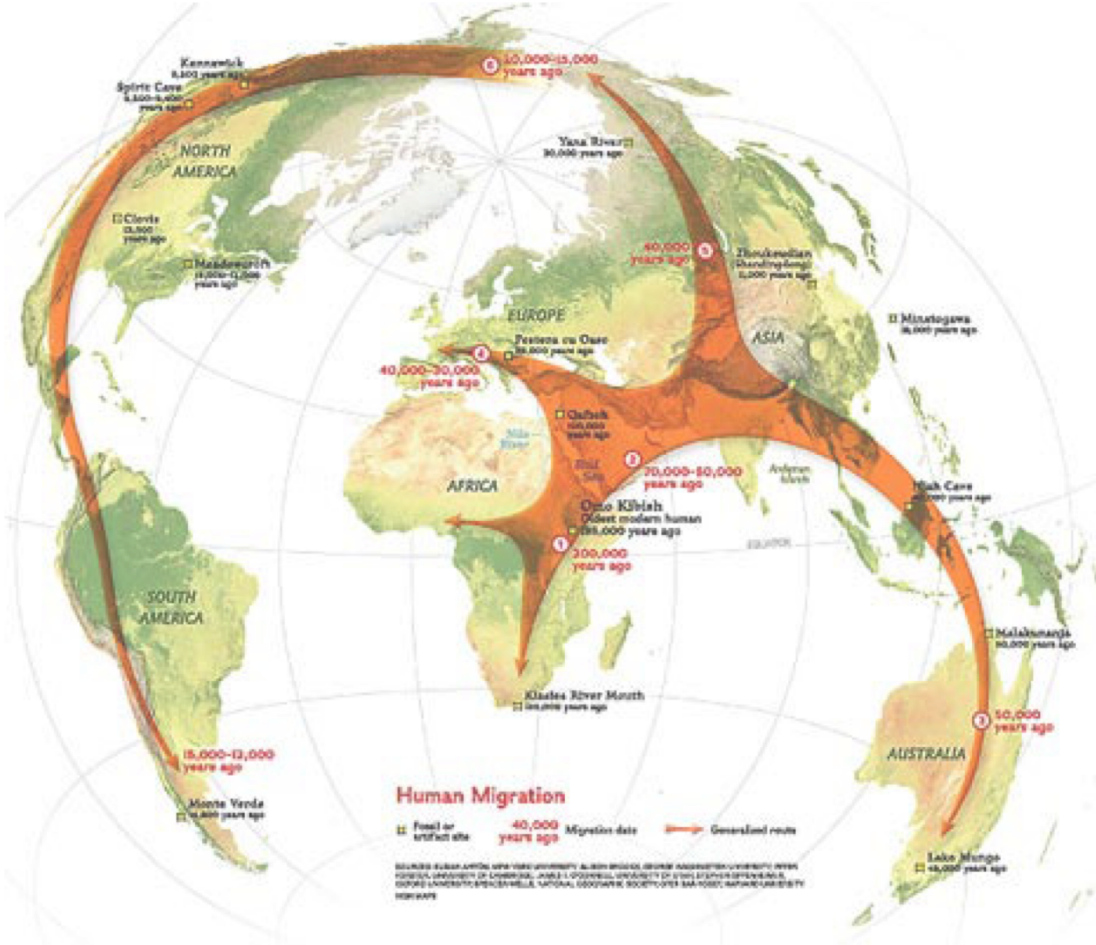
> If the lack of fit is bigger than expected → Reject H0 (the model)

> If the lack of fit is in the range → Do not reject H0

# Fitting data and measuring lack of fit is central to data analysis

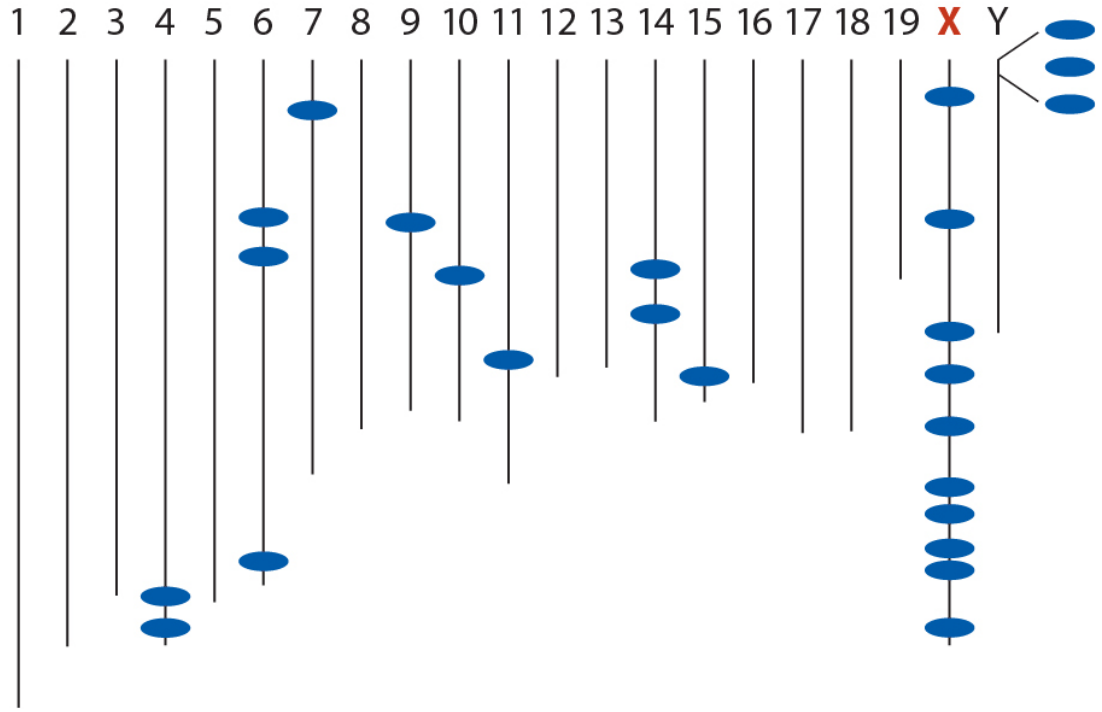



**Model** :
Genetic diversity decreases linearly away from East Africa**




** Because it is lost through sucessive genetic bottlenecks as modern humans expand ”out of Africa”

# We already met a goodness of fit test…



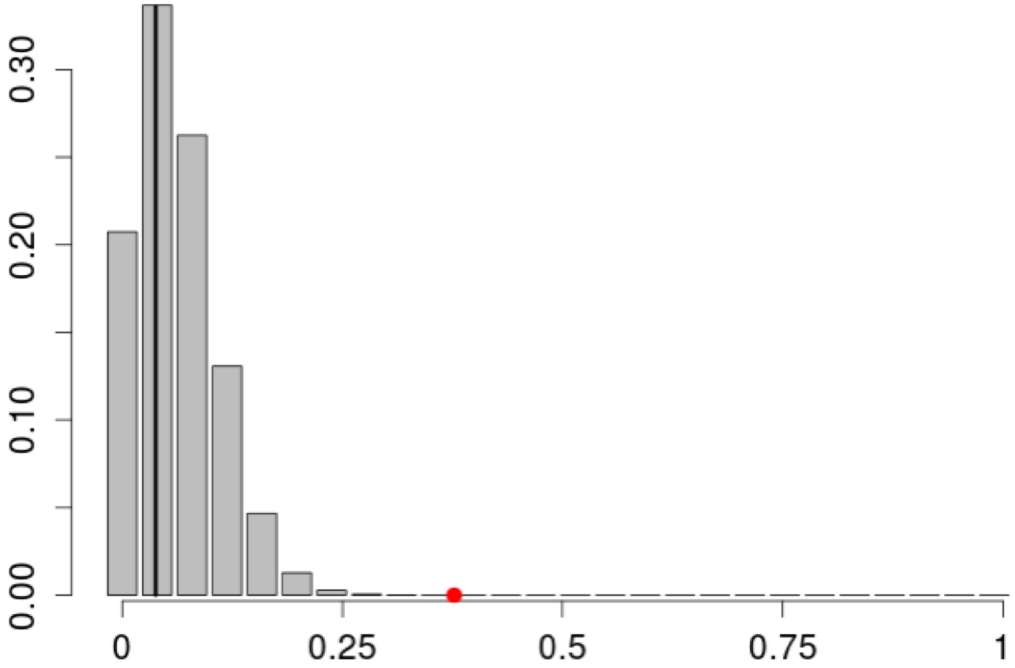

**Data:** 10 out 25 genes involved in spermatogenesis are on the X

H0: Proportional model with 2 categories

”**X chromsome**” 0.061

”**Rest of the genome**” 0.939

# We already did a goodness of fit test…




**Data:** 10 out 25 genes involved in spermatogenesis are on the X

H0: Proportional model with 2 categories

”X chromsome” 0.061

”Rest of the genome” 0.939

Test statistic: 10

Null distribution for the test statistic: binomial with n=25 trials and p= 0.061

# Generalizing GOF tests to several categories

**A (null) model can be mathematically derived or empirically based**

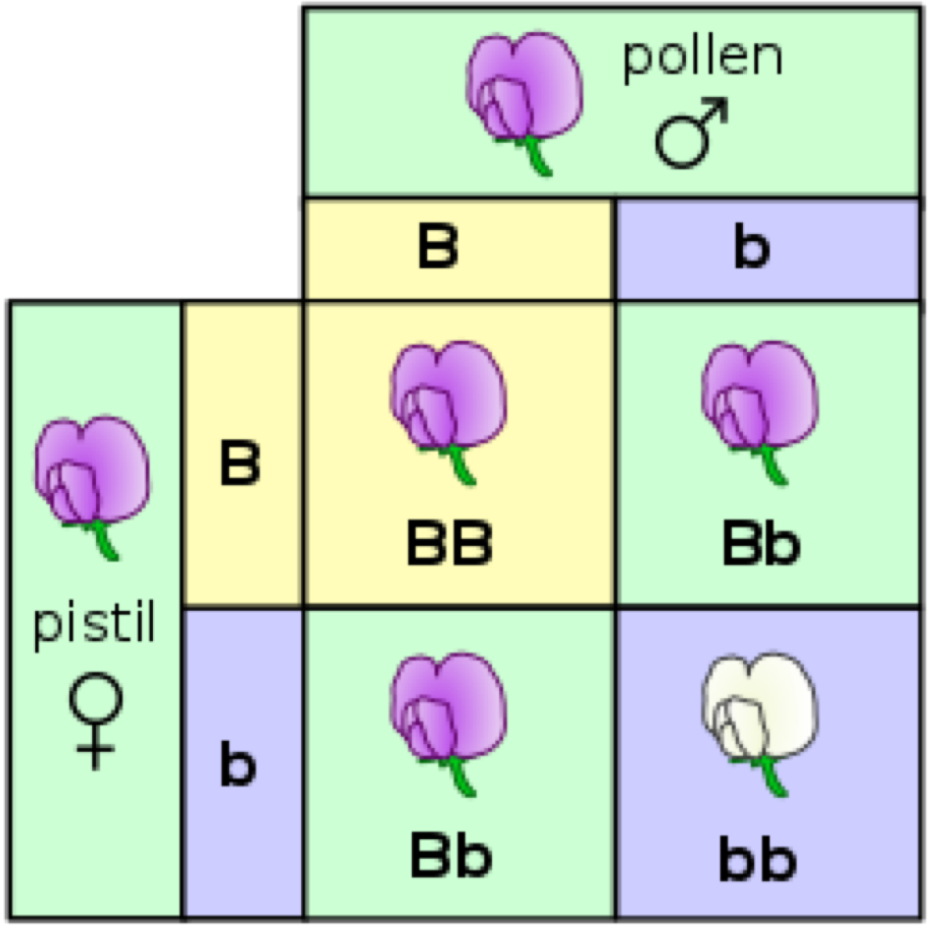
**It specifies what the data should look like**

**Mendel rules for segregation**: ¾ ¼

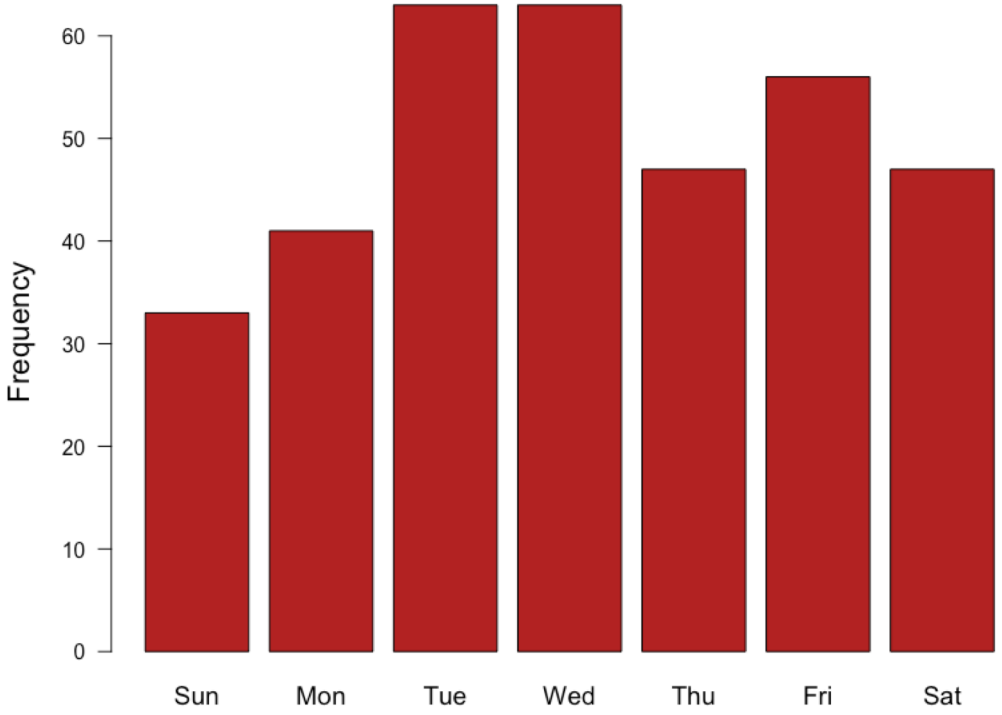
**Proportional (to equal opportunity)"model**

Spermatogenesis genes on the X

Birth on each week day



# Goodness-of-fit tests



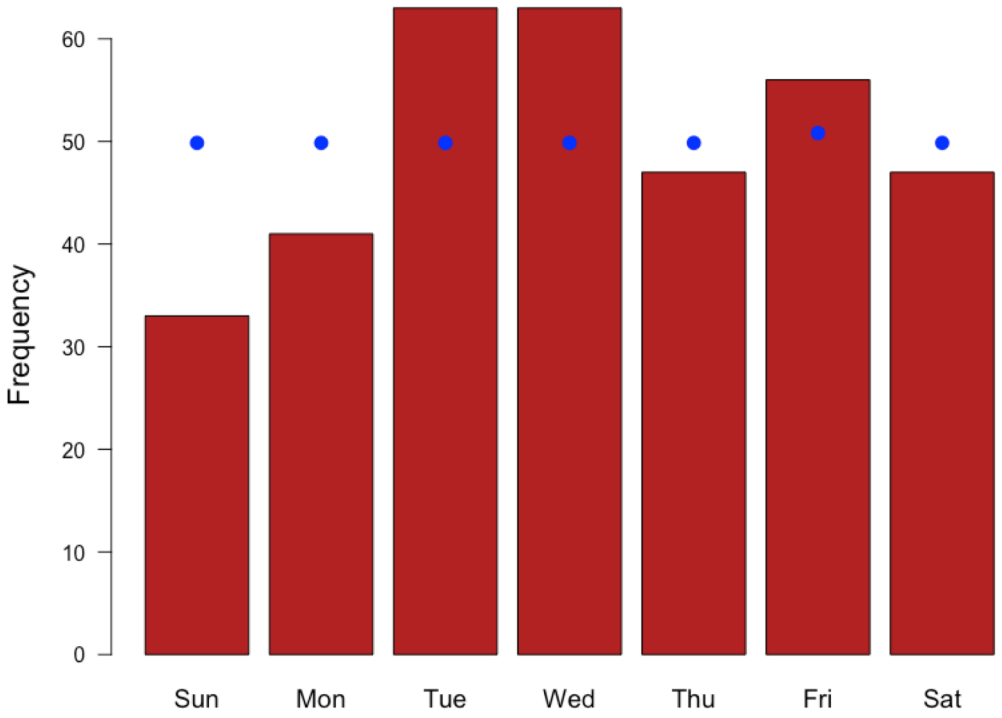

In a nutshell

We examine how well the observed data matches the expectations from a null model

→We summarize the observed data:

Observed counts in pre-defined categories

# Goodness-of-fit tests




In a nutshell

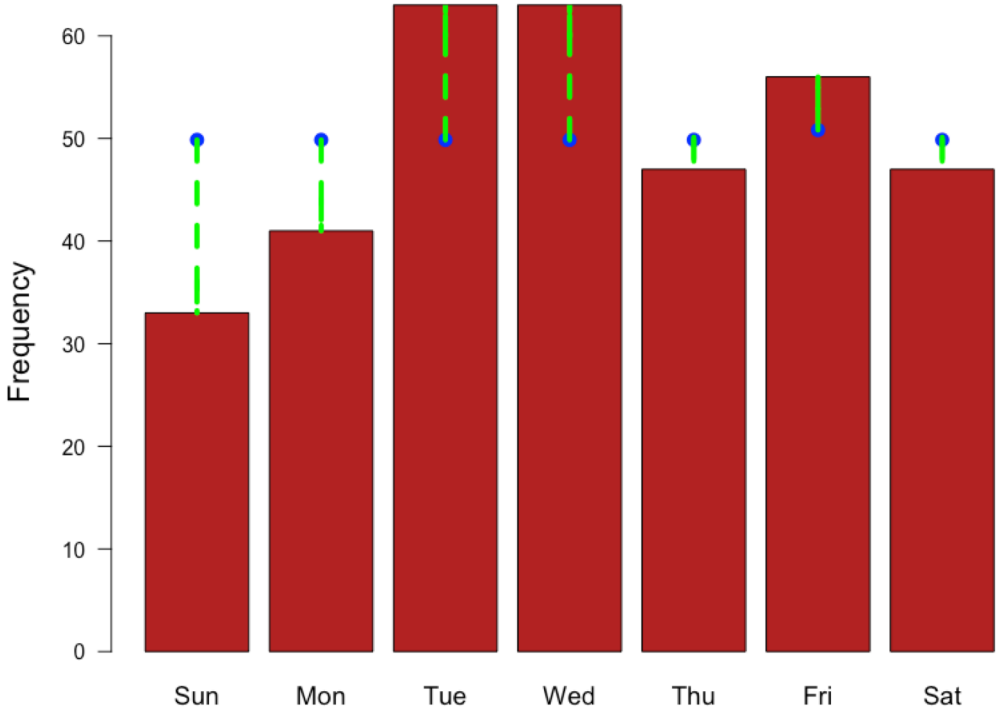
We examine how well the observed data matches the expectations under a null model

Here the null model is:

A proportional to ”opportunity” model

Expectations can be calculated BEFORE examining the data

# Goodness-of-fit tests



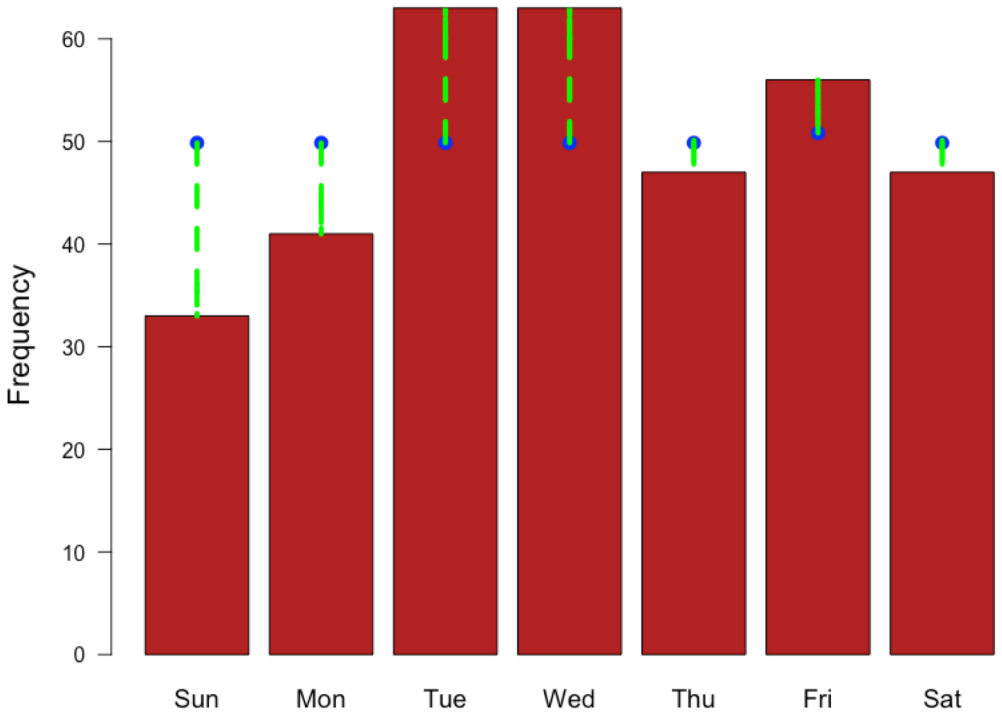

In a nutshell

We examine how well the observed data matches the expectations under our null model,

We quantify the **lack of fit**

$$\chi^2 = \Sigma_i (obs_i - exp_i)^2 / exp_i$$

# Goodness-of-fit tests: why the $\chi^2$ ??



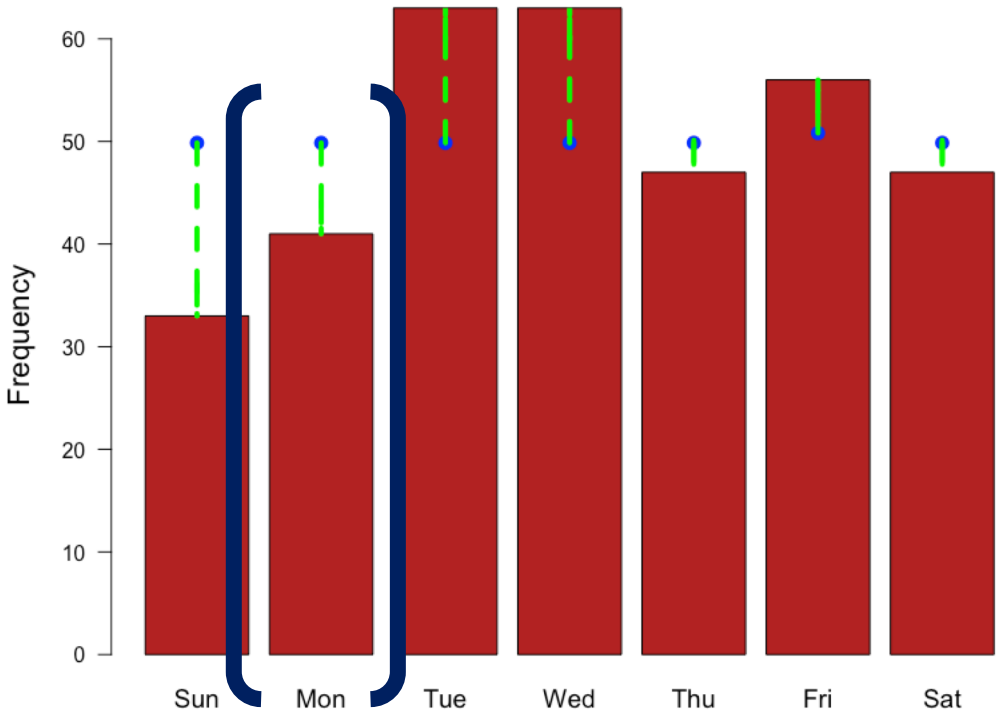

In GOF tests, we measure lack of fit using the ”almost holy” **test statistic**

$\chi^2 = \Sigma_i (obs_i - exp_i)^2 / exp_i$

That raises 2 questions:

- Why use $\Sigma_i (obs_i - exp_i)^2 / exp_i$ ?
- How much variation do we expect by chance alone in $\chi^2$ ?

# Decomposing the lack of fit: MONDAY




Let's start by focusing on one category: "Mondays"

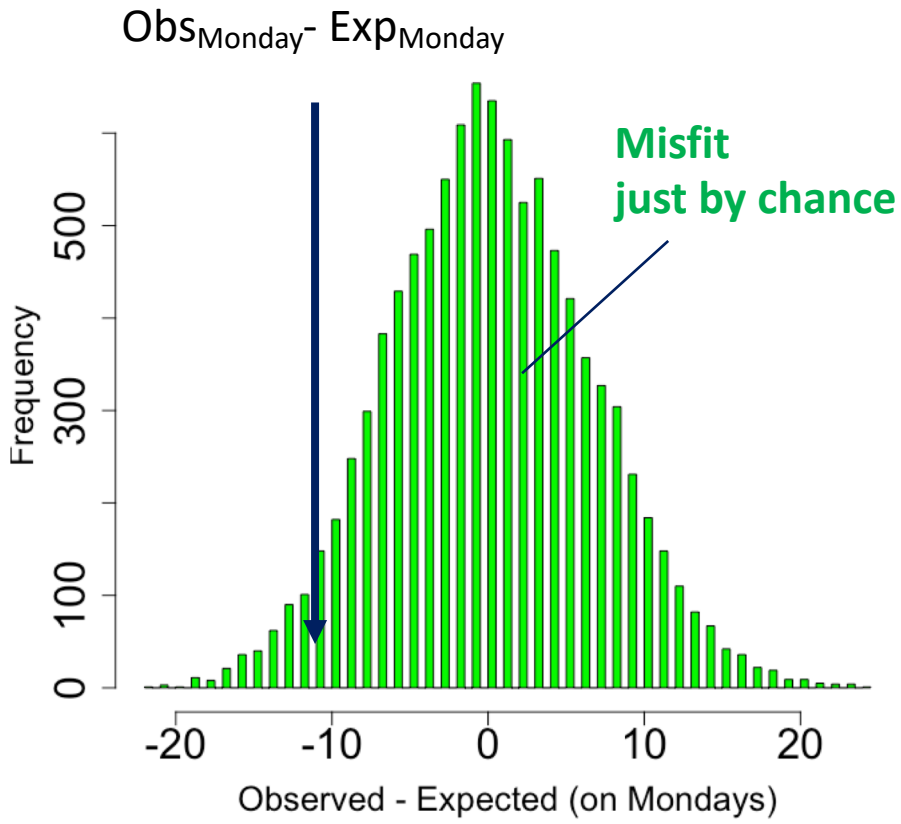
One natural simple measure is misfit= obs- expected

What can we say about misfit?

It is a random variable ..

The distribution of misfit is … ??

# Decomposing the measure of lack of fit




The simplest ”lack of fit” measure is misfit= obs- expected

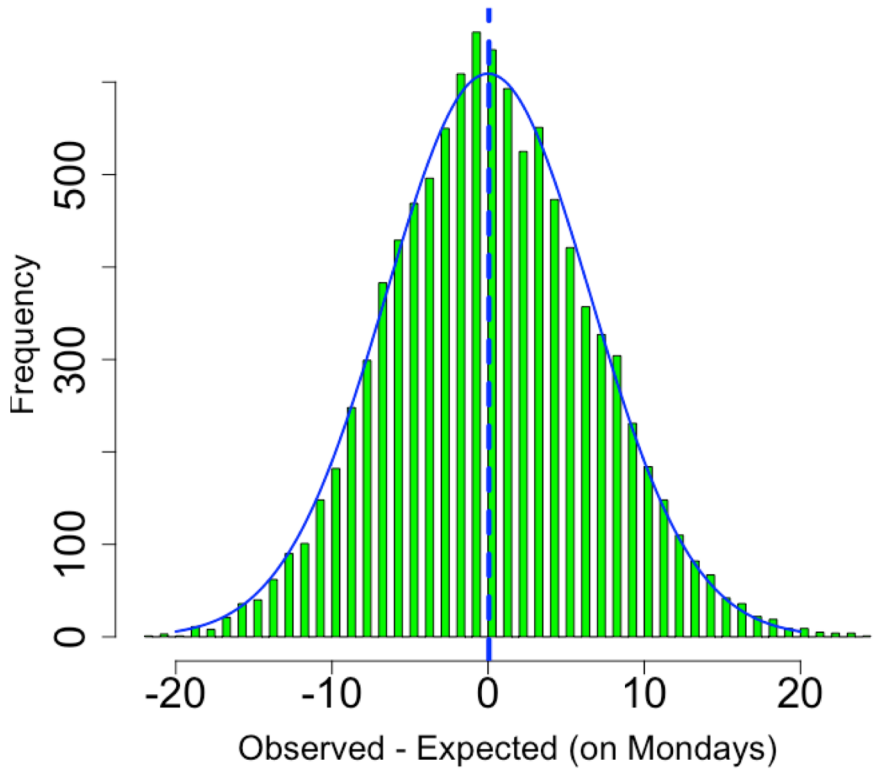
What can we say about misfit?

The **observed counts** for ”Mondays” is a random variable … and …

IT IS BINOMIALLY DISTRIBUTED ;-)

- n trials
- Fixed Probability of falling on ”Mondays”

←Here is a **simulation check**




So the misfit is like a ”shifted” binomial random variable

$E[\text{mistfit}_{Mon}] = E[\text{Obs}_{Mon}] - E[\text{Exp}_{Mon}]$

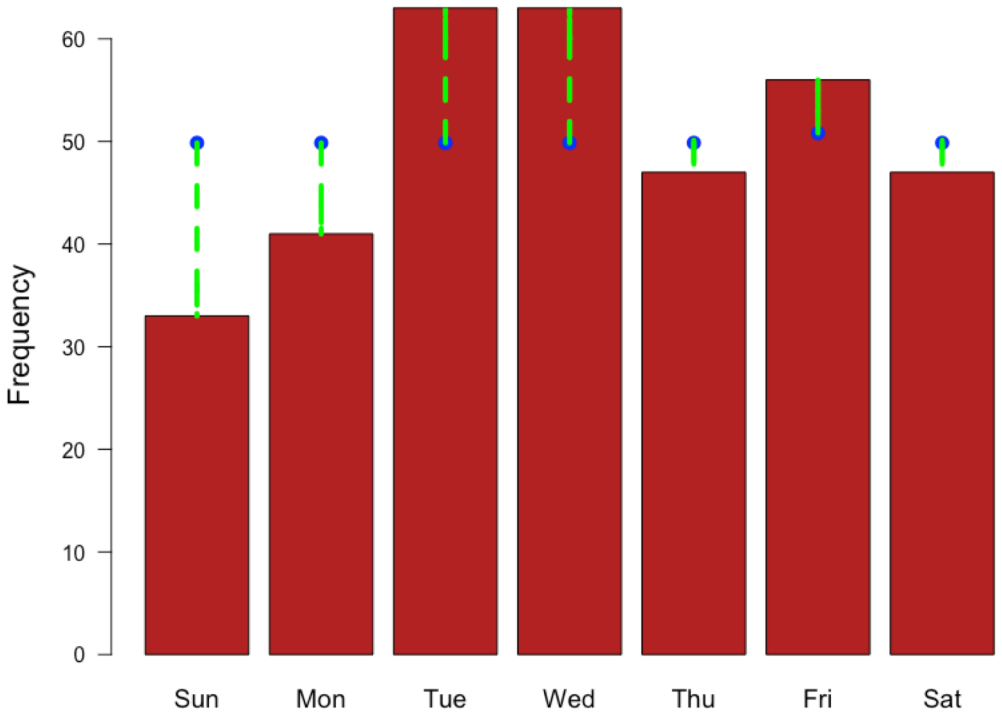
$E[\text{Obs}_{Mon}] = N\, p_{Mon}$ ($\text{Obs}_{Mon}$ is binomial)

$\text{Exp}_{Mondays} = N\, p_{Mondays}$ (We chose that)

$E[\text{mistfit}_{Mon}] = 0$

$\text{mistfit}_{Mon}$ **looks very bell-shaped** (normally distributed, see ch10)

# From misfit on ”Mondays” to overall lack of fit … rebuilding the $\chi^2$




Misfit on Tuesday , Wednesday is pretty much the same …

For an overall measure we want :

1. to avoid lack of fit on each day **cancelling out**

$$misfit_i \rightarrow misfit_i^{\,2}$$

2. Relative lack of fit

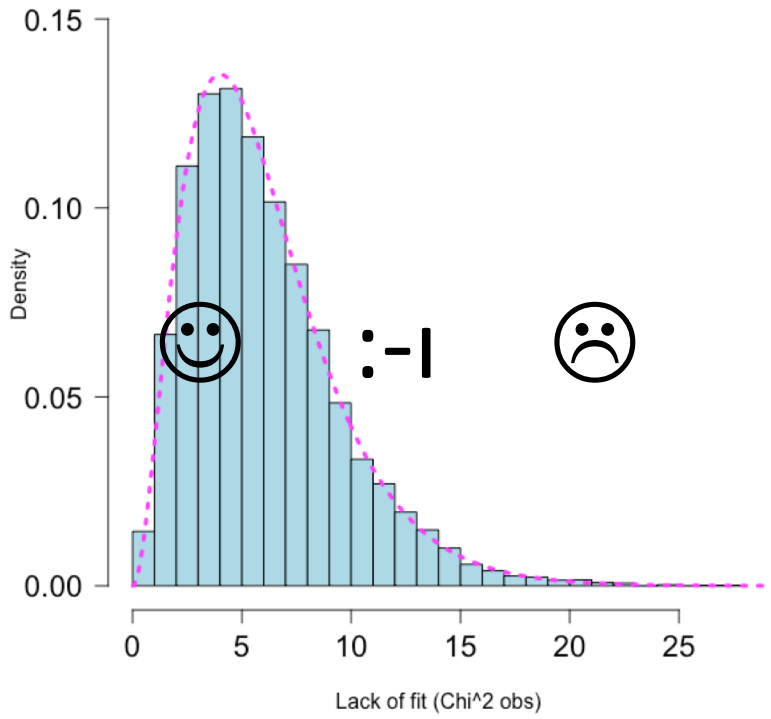
(so categories with large counts do not dominate the measure)

$$misfit_i^{\,2} \,/\, exp_i$$

**3. We addup the misfits terms:**

$$\chi^2 = \Sigma_i \, (obs_i - exp_i)^2 \,/ exp_i$$

# What lack of fit do we expect just by chance?




Simulations (Nsims= 10^3):

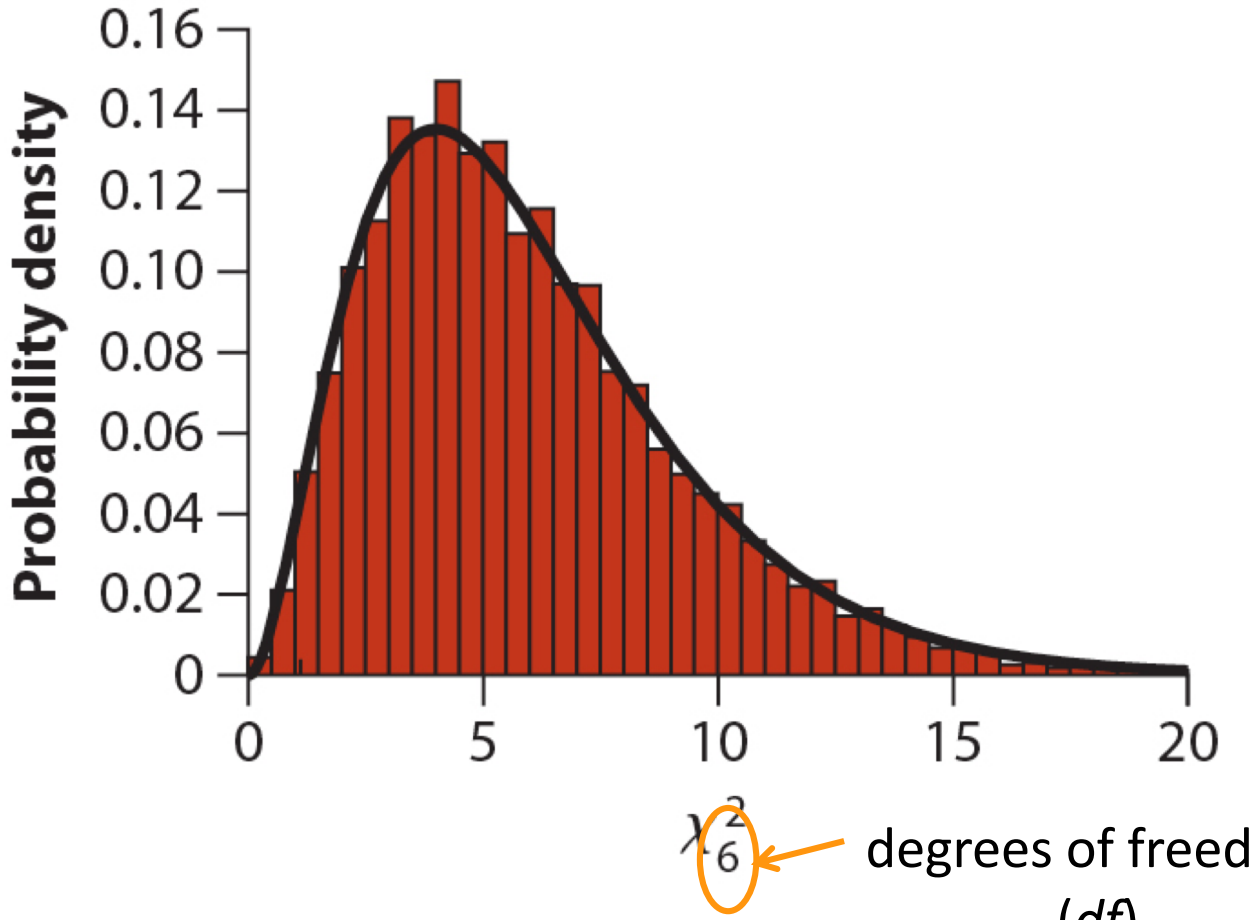
>Place 350 birthday at random according to the proportional model

Calculate the lack of fit ($\chi^2$)

There is a limiting probability distribution

→ ”Chi^2” approximation

# Sampling distribution of $\chi^2$ under $H_0$ : how many d.f. ?



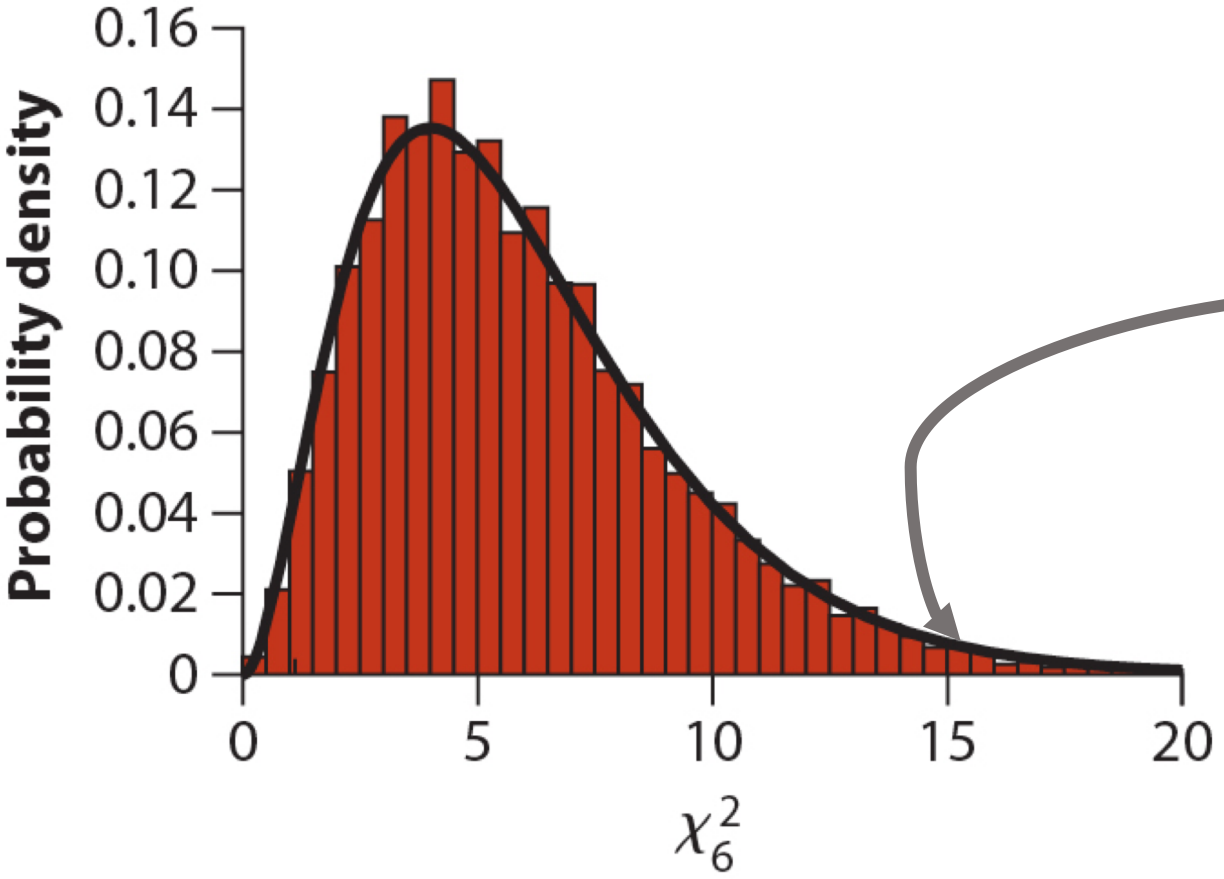

degrees of freedom (*df*)

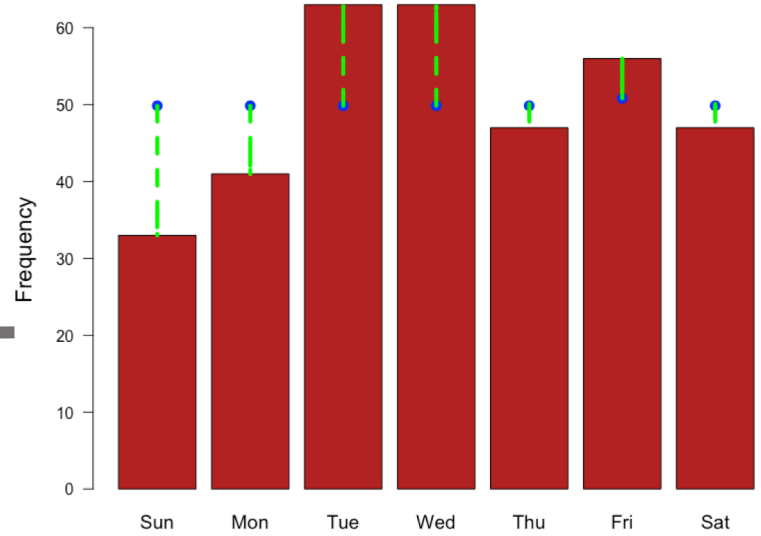
*df* = (number of categories) – 1 – (~~number of parameters estimated from data~~)

= 7 – 1 =6

Here **zero** for the proportional model

# Sampling distribution of $\chi^2_{obs}$ under $H_0$

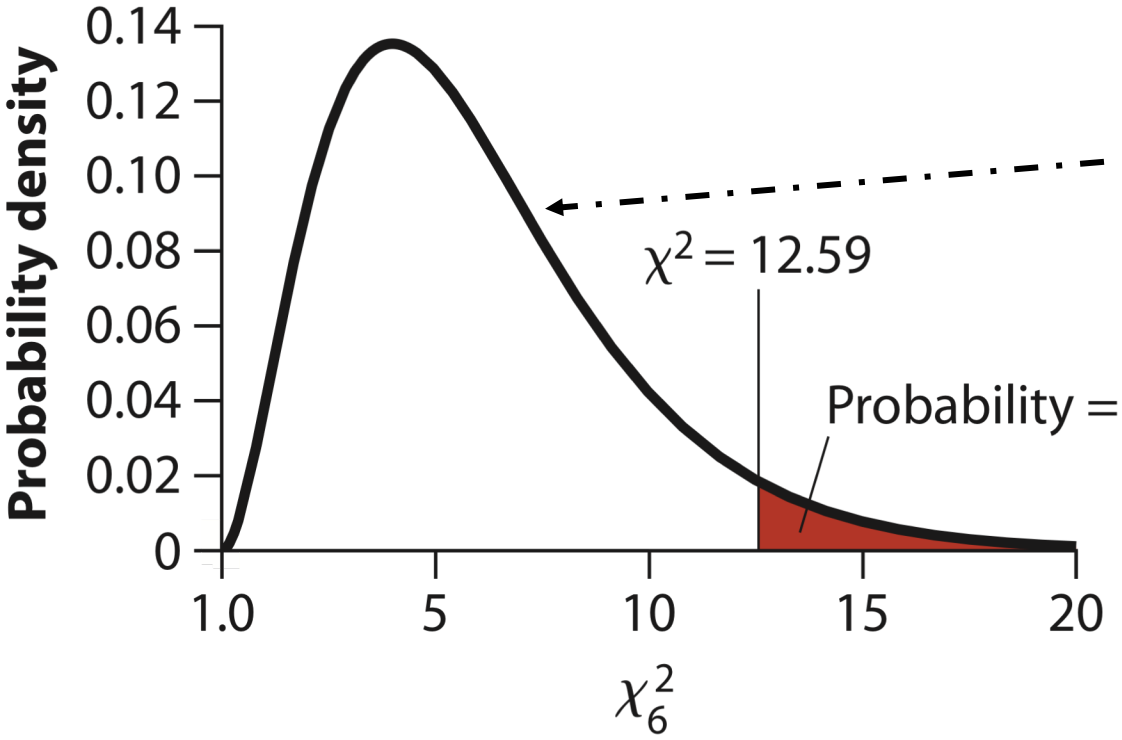



$\chi^2_{obs} = 15.05$

$P = 0.0199$

# A few things on $\chi^2$ probability distributions




A family of continuous distribution

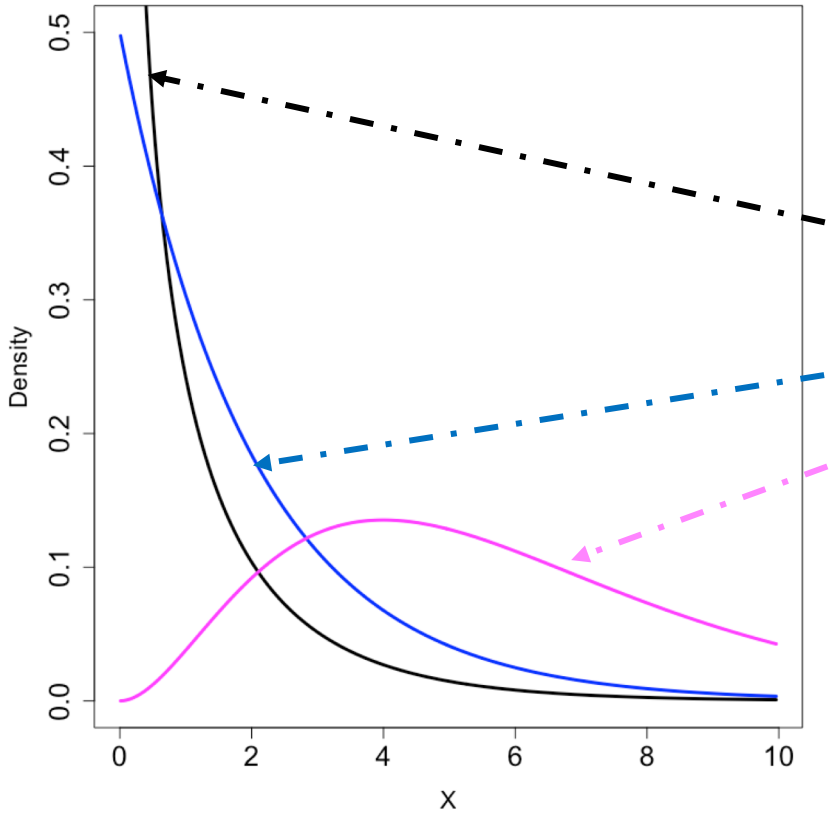
In R:

dchisq(...)

P(X > 12.59) = 0.05

In R 1- pchisq(...)

# on $\chi^2$ probability distributions
## Plotting densities in R




```
# xs: vector of coordinates on the X axis
xs= seq(from = 0.01, to=10, by=0.05 )


#y1 vector density of the Chi^2 (1 d.f)
y1=dchisq(x = xs , df = 1)
y2=dchisq(x =xs , df = 2)
y6=dchisq(x =xs , df = 6)
plot(xs,y1, type ="l", lwd=3, xlab = "X",
ylab="Density", cex.axis=1.6, cex.lab=1.4,
ylim=c(0,.5))
lines(xs,y6, type="l", lwd=3,
col="magenta") es(xs,y2, type ="l", lwd=3,
col="blue")
```

# ***RECAP: GOF testing with $\chi^2$ ***

A. Formulate $H_0$
B. Get the expected counts under $H_0$
C. Calculate the observed lack of fit
   > test statistic $\chi^2_{obs}$ from the data

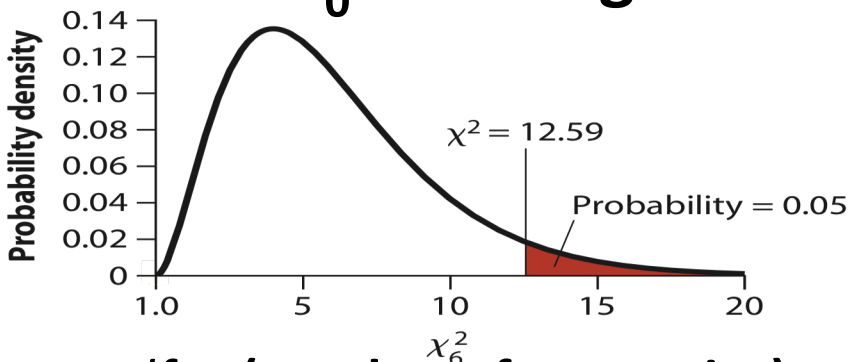
D. Figure out the distribution of $\chi^2_{obs}$ under $H_0$
   **Limiting distribution** ($\chi^2$)
   **Simulated distribution**

E. Get a p-value & take a decision

P-value= $P[\chi^2_{df} > \chi^2_{obs}]$

P-value > $\alpha$ fail to reject $H_0$
P-value < $\alpha$ reject $H_0$

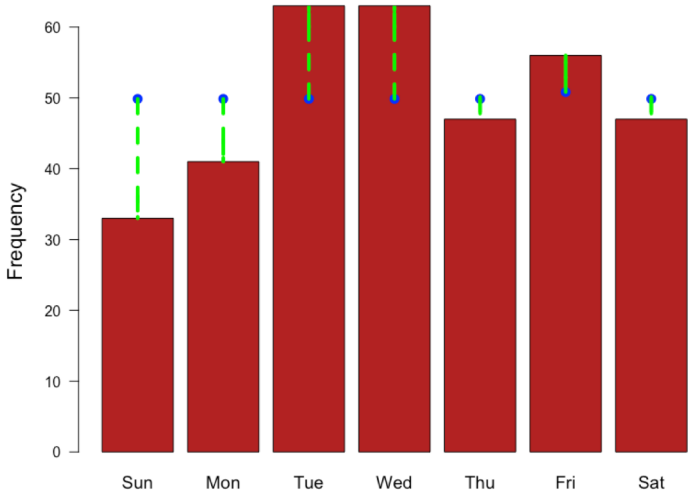
**$H_0$ "limiting" distribution**

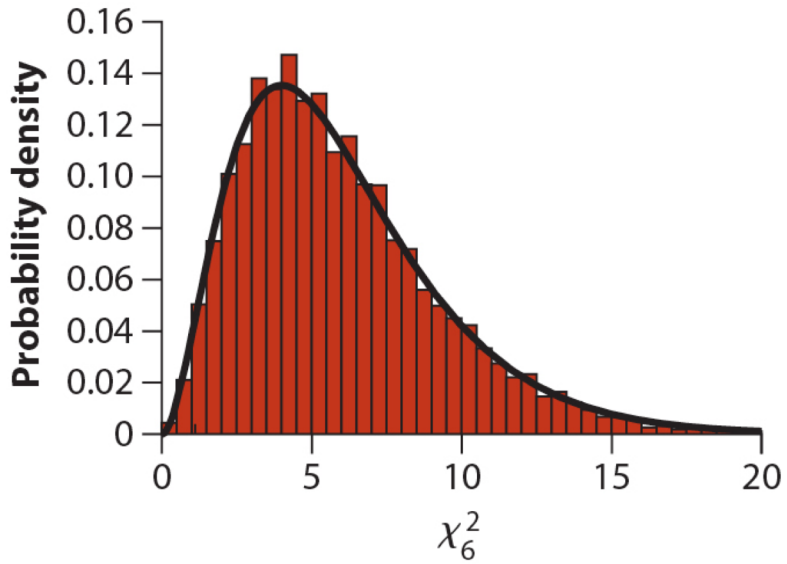



*df* = **(number of categories) – 1**

– (number of parameters estimated from data)

# Assumptions of $\chi^2$ goodness-of-fit test




The limiting distribution for H0 works well as long as the following rules are obeyed:

No categories have expected frequency less than 1

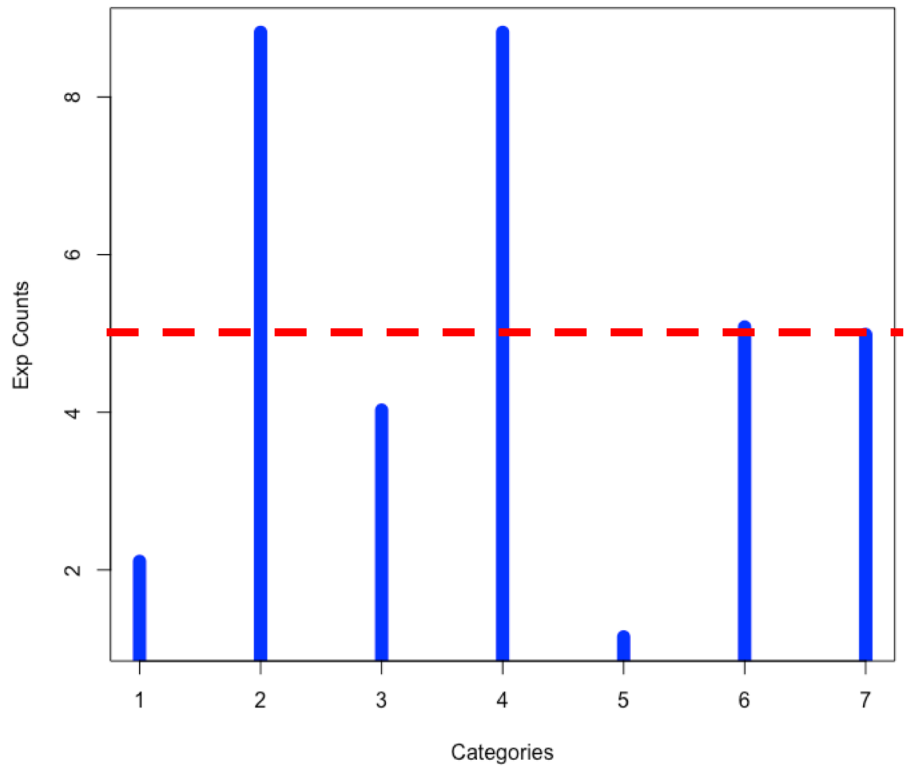
No more than 20% of categories have expected frequencies less than 5

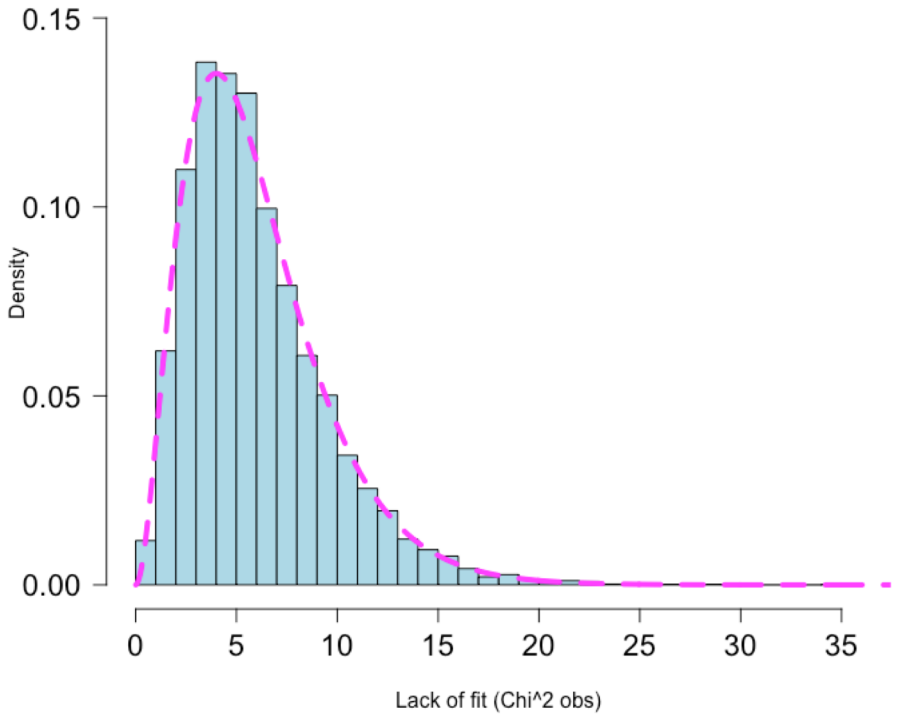
If not :

POOL categories (ancient)

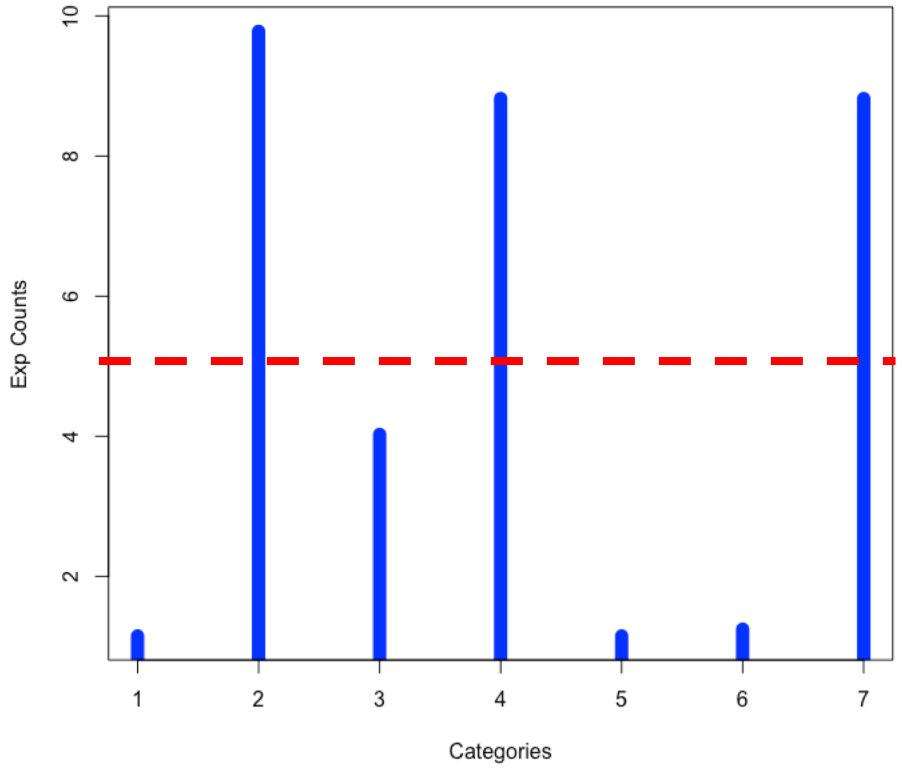
SIMULATE H0 (modern)

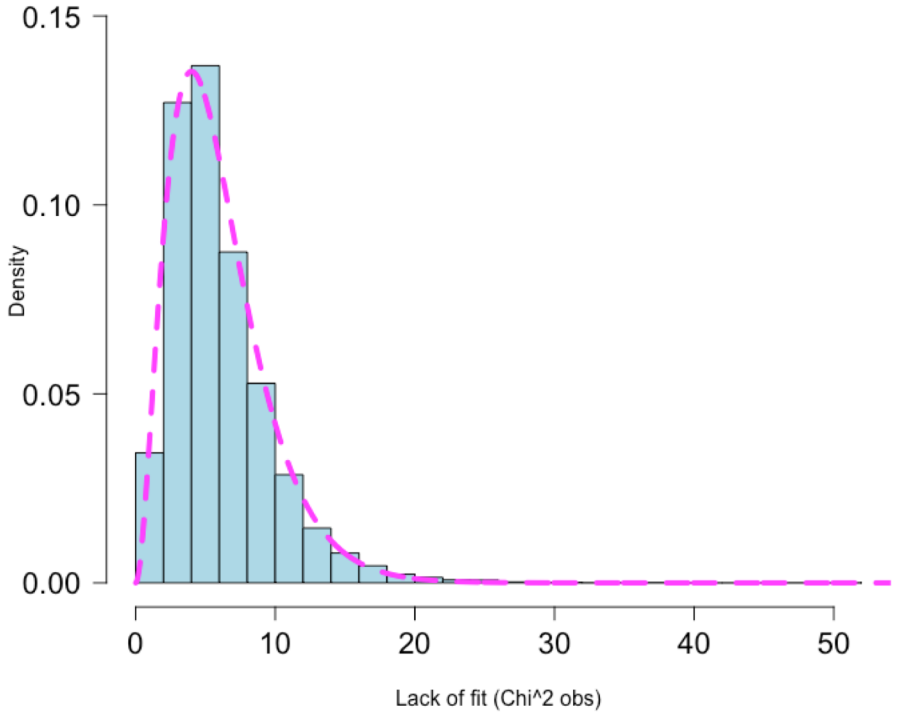
# The limiting distribution is very robust n=35 with unbalanced $H_0$



# The limiting distribution is quite robust n=35 with very unbalanced $H_0$



# A bit about the exercise of Thursday

Mendel 84 experiments

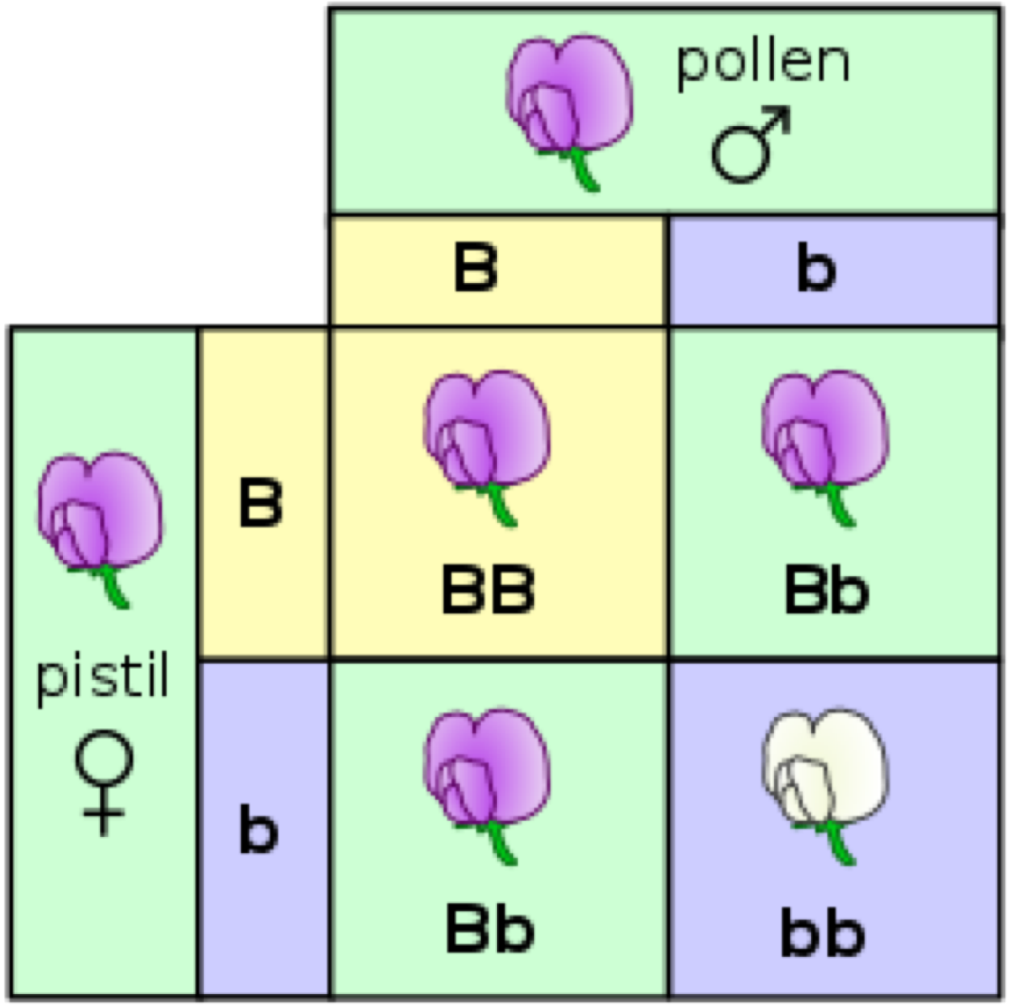
Fitting a binomial distribution to
count data

A recap on GOF tests

# Mendel Laws of inheritance: a recap …

Cross two true breeding lines …
and self the F1

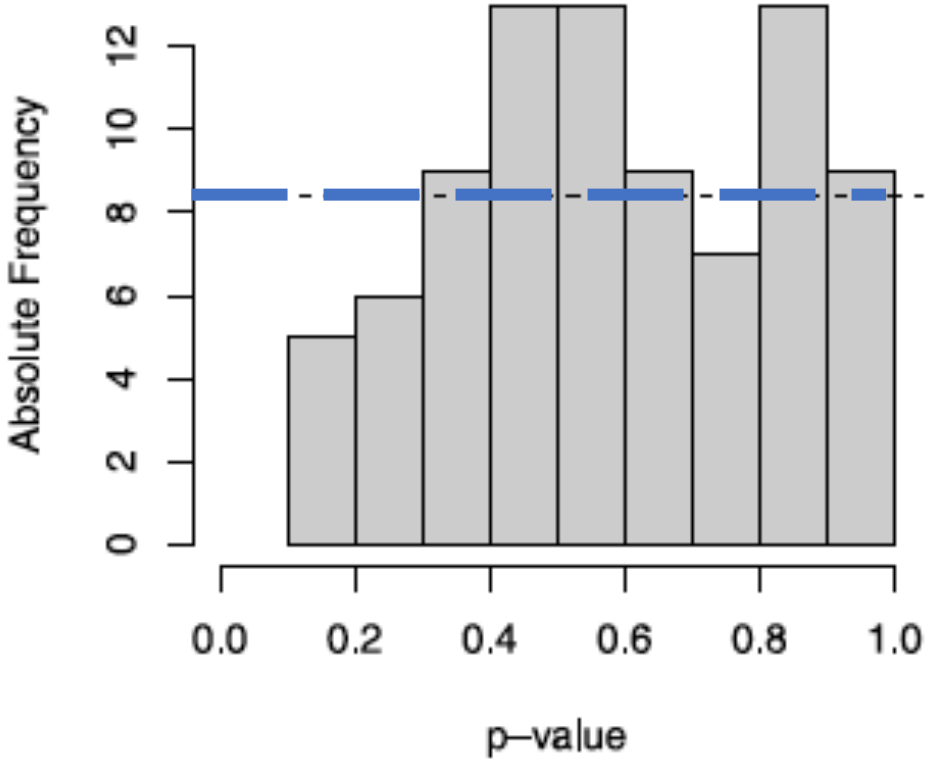
H0: F2 we expect 2 categories of outcomes in proportion ¾ and ¼



# A snapshot of the (BIG) data

| | Trait | "A" | "a" | $n$ | Obs. freq. $n$"A" | Obs. freq. $n$"a" | Theor. ratio "A" : "a" |
|---|---|---|---|---|---|---|---|
| | Seed shape | round | wrinkled | 7324 | 5474 | 1850 | 3 : 1 |
| | Seed color | yellow | green | 8023 | 6022 | 2001 | 3 : 1 |
| | Flower color | purple | white | 929 | 705 | 224 | 3 : 1 |
| $F_2$ | Pod shape | inflated | constricted | 1181 | 882 | 299 | 3 : 1 |
| | Pod color | yellow | green | 580 | 428 | 152 | 3 : 1 |
| | Flower position | axial | terminal | 858 | 651 | 207 | 3 : 1 |
| | Stem length | long | short | 1064 | 787 | 277 | 3 : 1 |

# Looking at all 84 experiments by Mendel

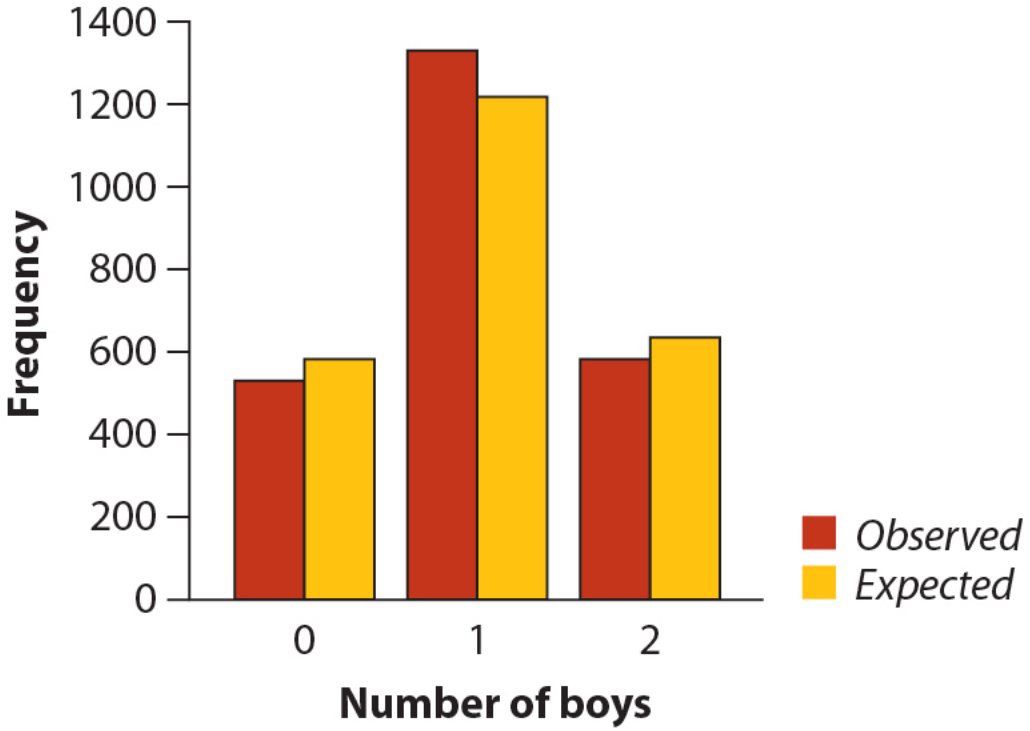



Can the fit of a model to the data be ”too good”?

Important fact (more about that after Easter with Palle):

If all datasets come from H0, we expect ?

# Fitting a binomial distribution to data (1/2)




Families of 2 kids (n=2444 families)

| DATA | 0 | 1 | 2 |
|---|---|---|---|
| obs | 530 | 1332 | 582 |

$H_0$ **Number of Boys per family** is binomial

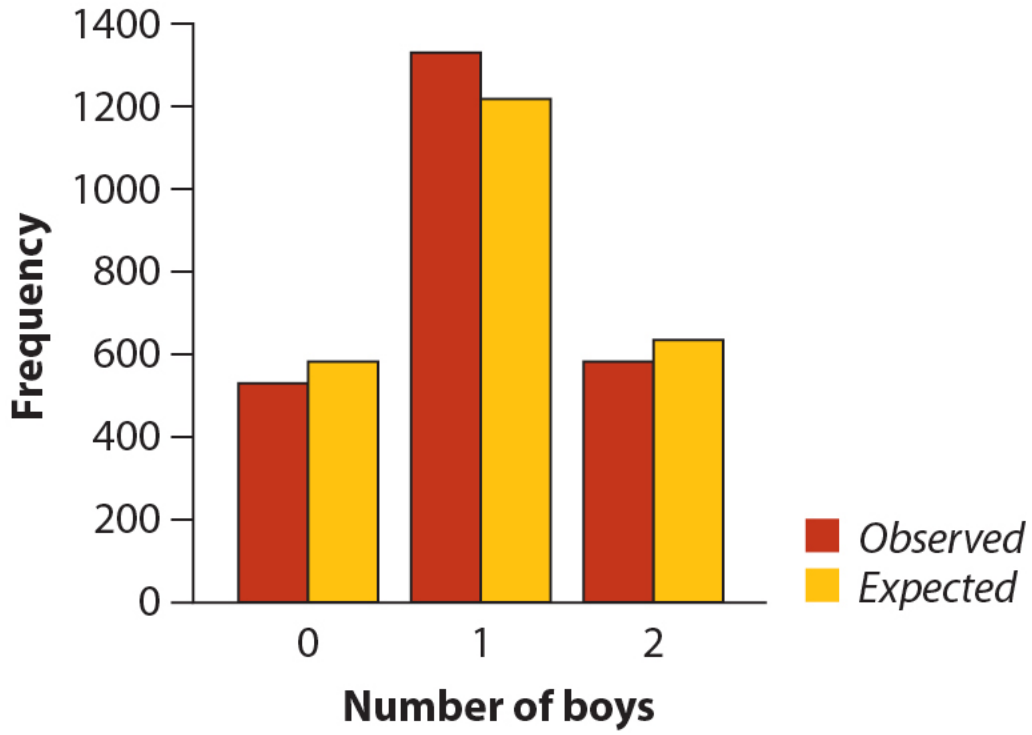
$H_A$ Number of Boys is NOT binomial

**Under H0 is formulated, we cannot specify $p_{boy}$ BEFORE we fit the data**

$N_{boy}$= 2*582 + 1332 =2496

Overall $p_{boy} = N_{boy}/ 2*n$ =0.5106

# Fitting a binomial distribution to data (2/3)




$H_0$ Number of Boys in 1 family

is binomial (n= 2 , p= 0.5106)

P(0 boy), P(1 boy) P(2 boys) calculated

Expected number per class

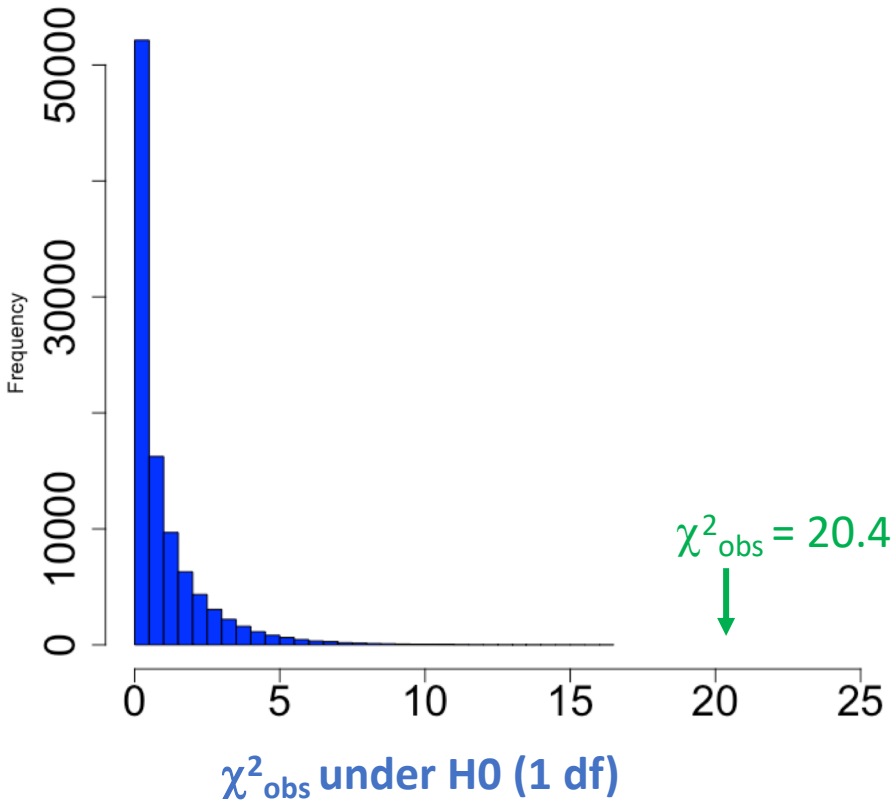
P(0 boy) * 2444 =585.3

P(1 boy) * 2444 = 1221.4

P(2 boy) * 2444=637.3

$\chi^2_{obs} = 20.4$

# Fitting a binomial distribution to data (3/3)




$\chi^2_{obs}$ = 20.4 is to be compared with the limiting a distribution under H0

Pvalue is in R

```
>1-pchisq(q = 20.3,df=1)
>6.620058e-06
```

Clearly the observed lack of fit is very unlikely under our null

Note the ”qualitative ” fit is not BAD but we have many many families …

# RECAP 1: Essential steps GOF testing with $\chi^2$

A. Formulate $H_0$
B. Get the expected counts under $H_0$
C. Calculate the observed lack of fit
   > test statistic $\chi^2_{obs}$ from the data

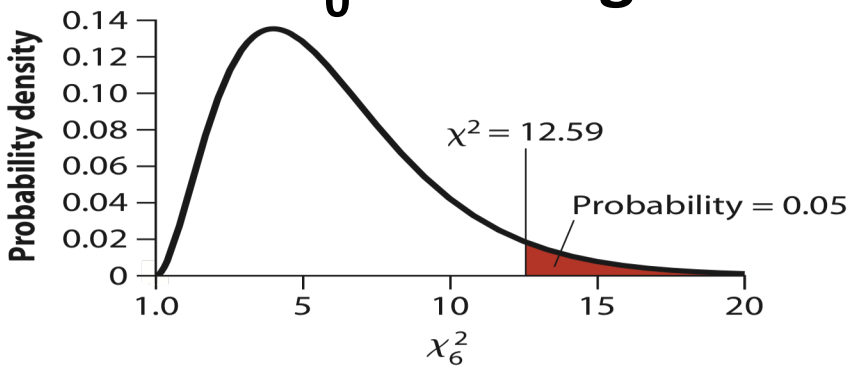
D. Figure out the distribution of $\chi^2_{obs}$ under $H_0$
   **Limiting distribution** ($\chi^2$)
   **Simulated distribution**

E. Get a p-value & take a decision

P-value= $P[\chi^2_{df} > \chi^2_{obs}]$

P-value > $\alpha$ fail to reject $H_0$
P-value < $\alpha$ reject $H_0$

**$H_0$ "limiting" distribution**




*df* = **(number of categories) – 1**
– (number of parameters estimated from data)

# Recap 2
# Fitting data to a null model: 2 different cases

**A specific** Proportional model **(8.1 & 8.4)**

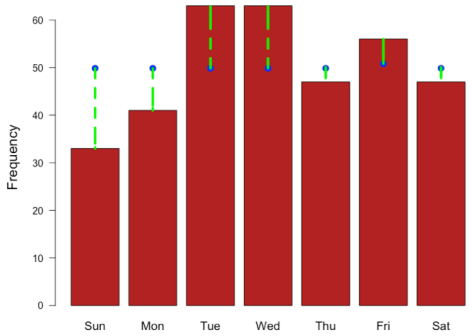
(**specified a priori**)

The expected count in each class is …

Calculate the lack of fit ($\chi^2_{obs}$ )

Compare with the null distribution for $\chi^2_{obs}$

*df* = **number of categories - 1**




A binomial distribution **(8.5)**

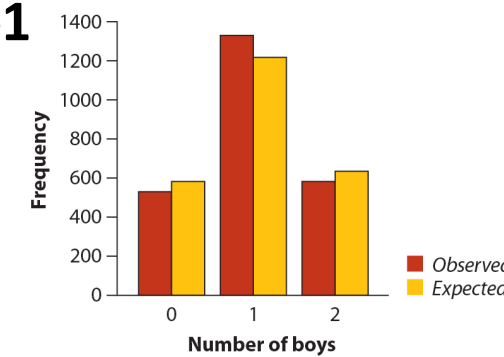
(**specified a posteriori**)

Choose the best binomial distribution (**fit p**) and calculate expected counts using this binomial

Calculate the lack of fit ($\chi^2_{obs}$ )

Compare with the null distribution for $\chi^2_{obs}$

*df* = **number of categories - 1 -1**



# 3 fundamentals question to classify ”random”?

**Q1: What happened: this or that ?**

Flip a coin, roll a dice

Genotype an individual

Classify a gene

Boy or girl ?

Birth on Monday or rest of the week

→ **Binomial** distribution N trials p

**Q2: How long do I have to wait before something happens ?**

**→Geometric distribution**

**P(waiting t times)=**

**$(1-p)\ (1-p)\ \ldots(1-p)\ p = (1-p)^{t-1}\ p$**

**Q3: How many things happened per unit time/space?**

**RELEVANT FOR :**

How many bacteria colonies are resistant to a virus/antibiotic?

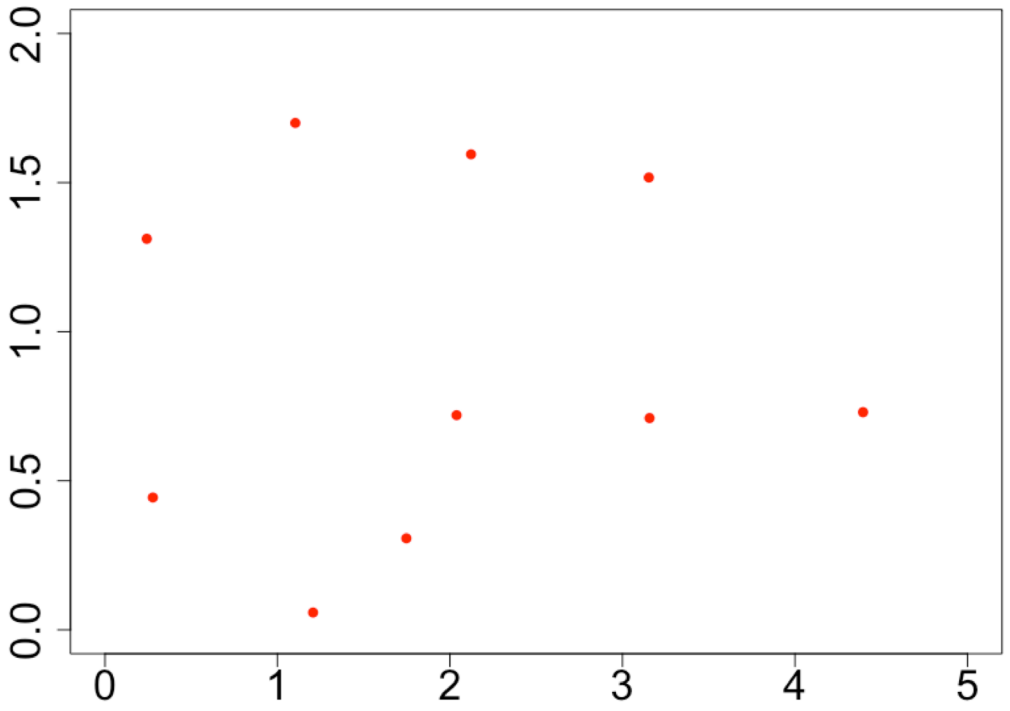
How many cases of a rare disease recorded in 2017 per 10,000 inhabitants ?

How many mutations accumulated in the human genome in 1or 100 generations?

How many sequences are ”covering” a genome region of interest ?

Are proteins evolving according to a molecular clock ?

# Random in space: 10 events in 10 $m^2$



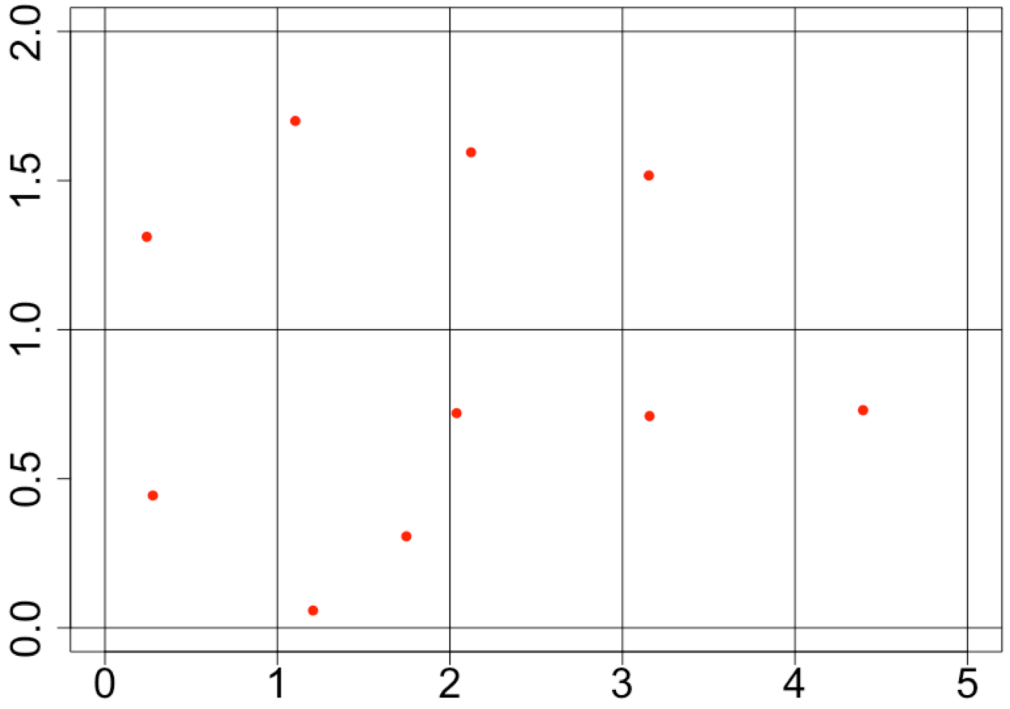

A surface of 10 square meters (5m x 2)

Random events:

(Spider falling from the sky) happen with equal (uniform) probability

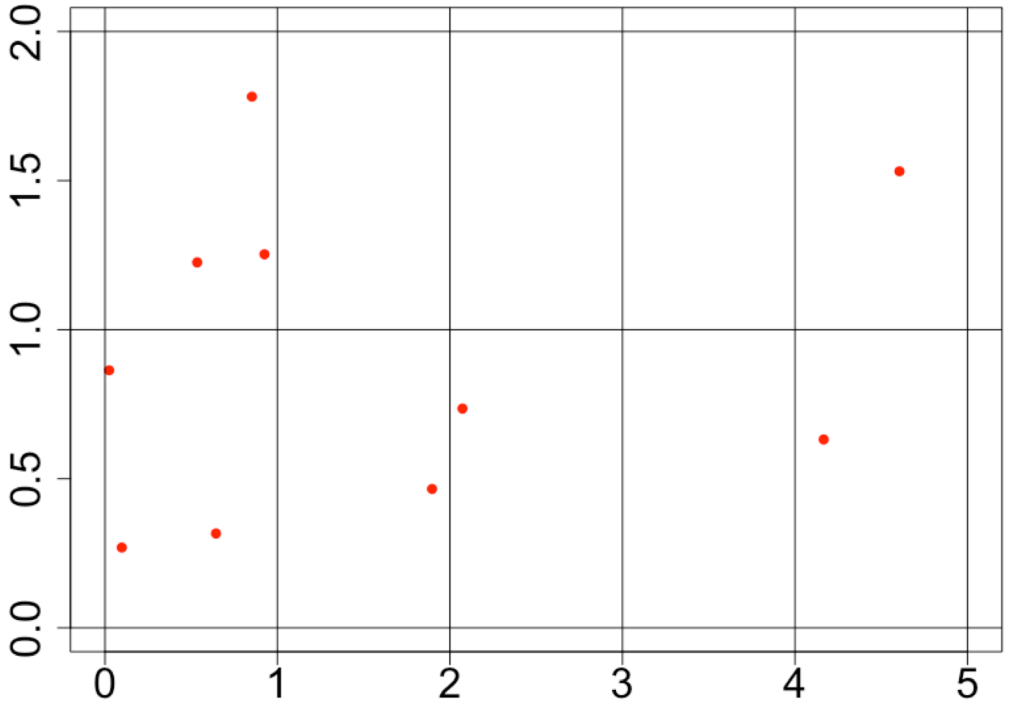
Events are ”blind to each other” (no spatial correlation)

# Counting events in units of $1m^2$



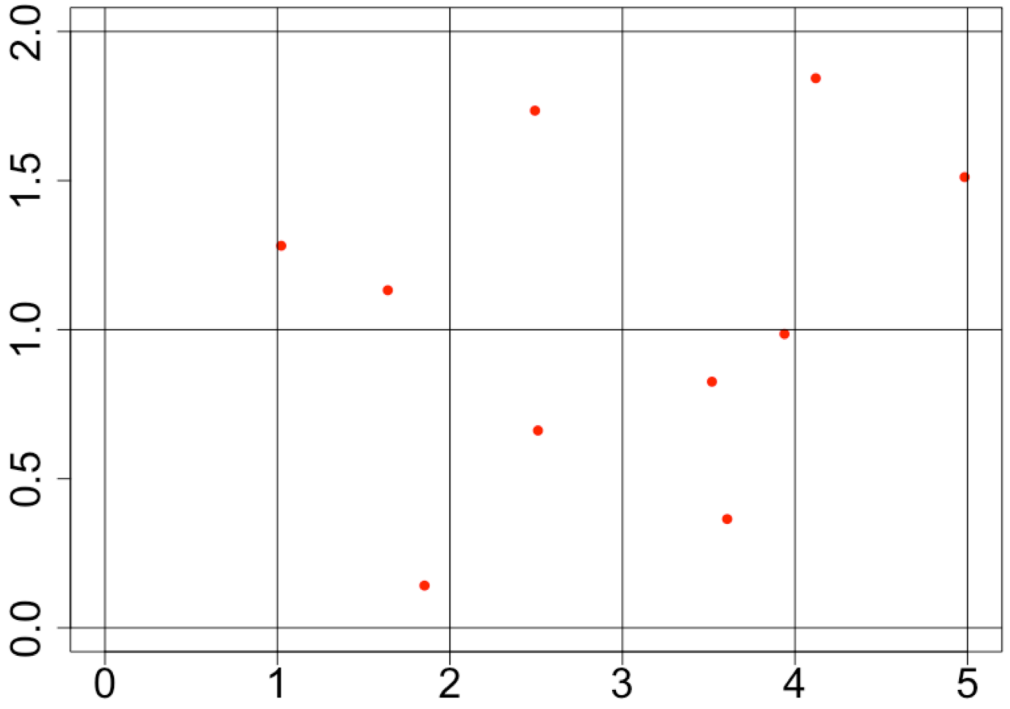

| Number Events | Frequency |
|---|---|
| 0 | 1 |
| 1 | 8 |
| 2 | 1 |

# Counting events in units of $1m^2$



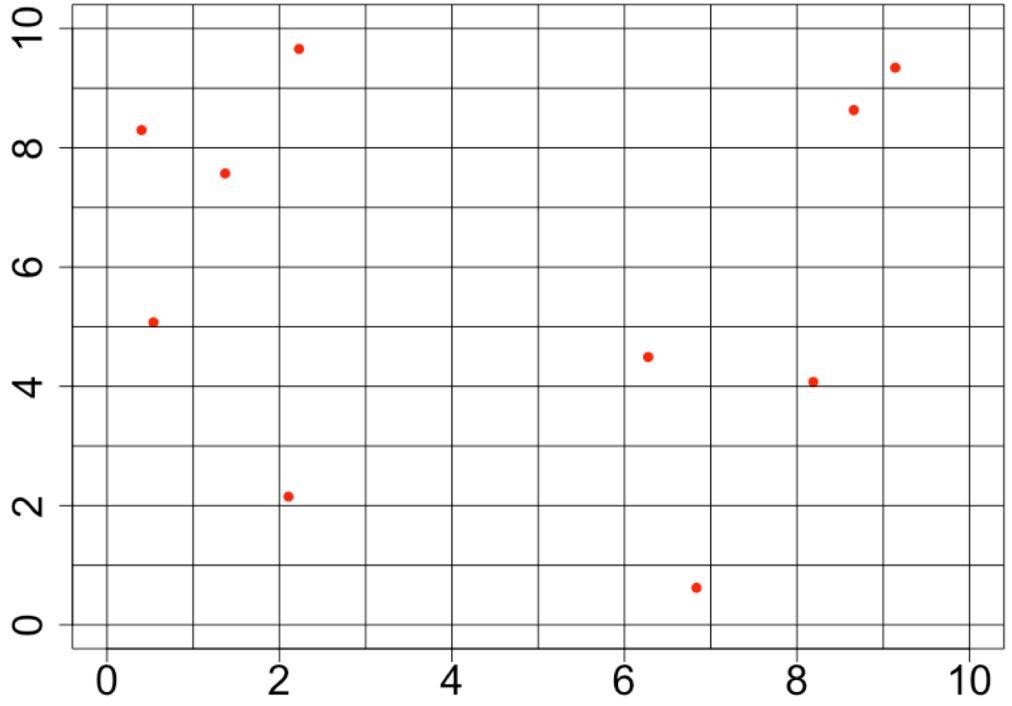

| Number Events | Frequency |
|---|---|
| 0 | 4 |
| 1 | 4 |
| 2 | 0 |
| 3 | 2 |

# Counting events in units of $1m^2$



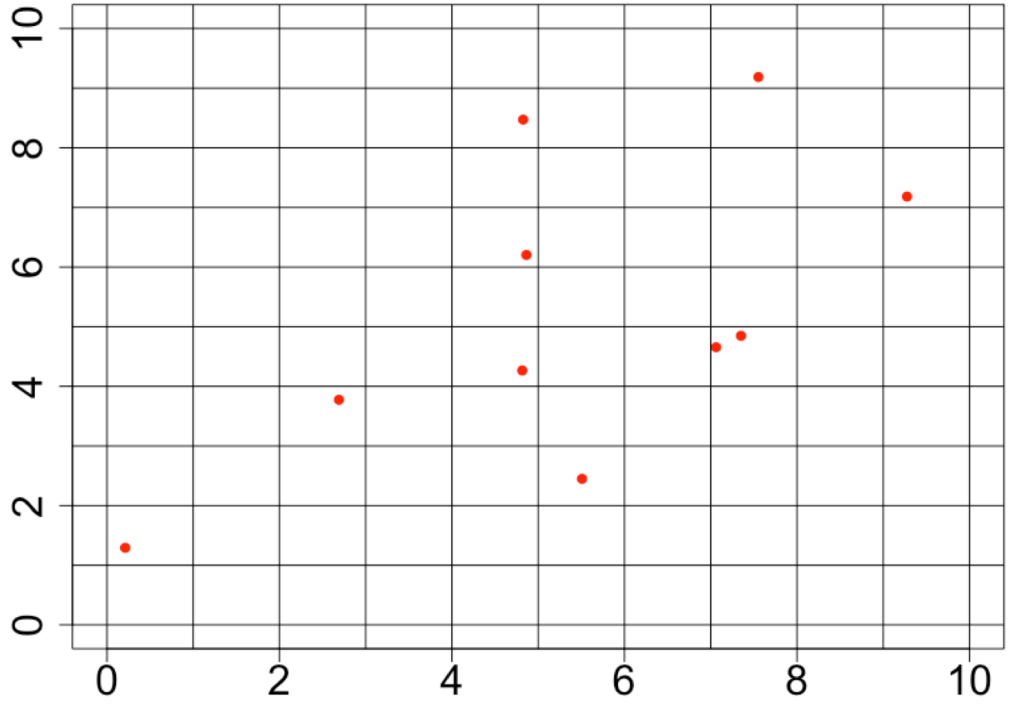

| Number Events | Frequency |
|---|---|
| 0 | 4 |
| 1 | 3 |
| 2 | 2 |
| 3 | 1 |

# 10 events in 100 $m^2$ → 0.1 event per unit



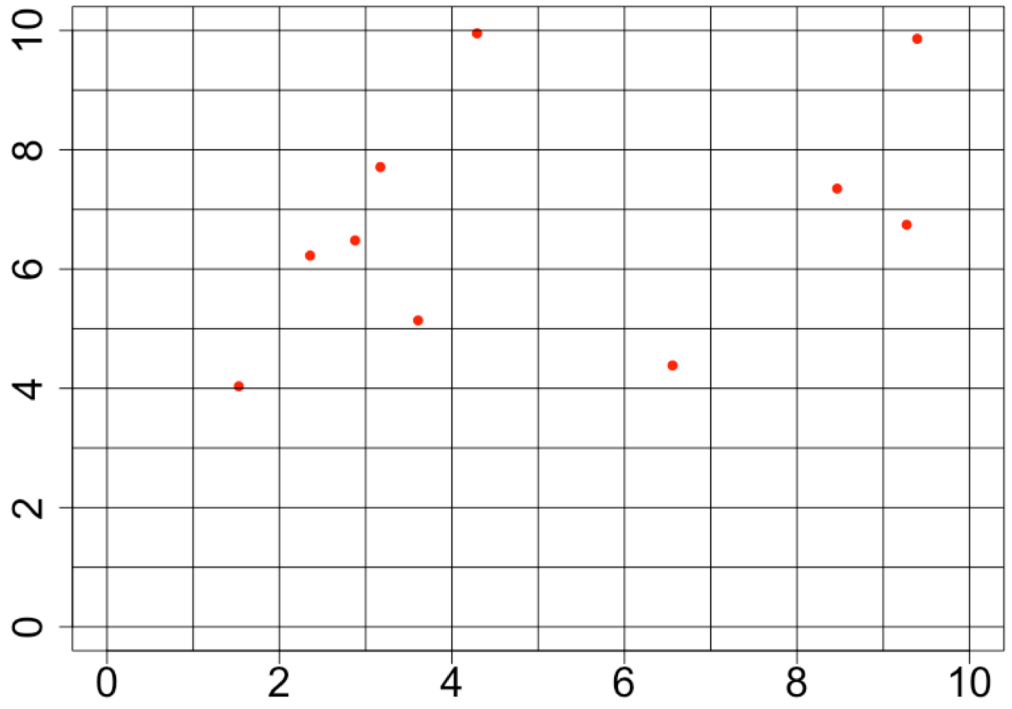

| Number Events | Frequency |
|---|---|
| 0 | 90 |
| 1 | 10 |

# 10 events in 100 $m^2$




| Number Events | Frequency |
|---|---|
| 0 | 90 |
| 1 | 9 |
| 2 | 1 |

# Doing it again and again (10 times)




**10 times**

I have surveyed **1000 units of 1m²**

**On average** 0.1 events per m²

| Events /m² | 0 | 1 | 2 |
|---|---|---|---|
| Number of obs | 903 | 94 | 3 |

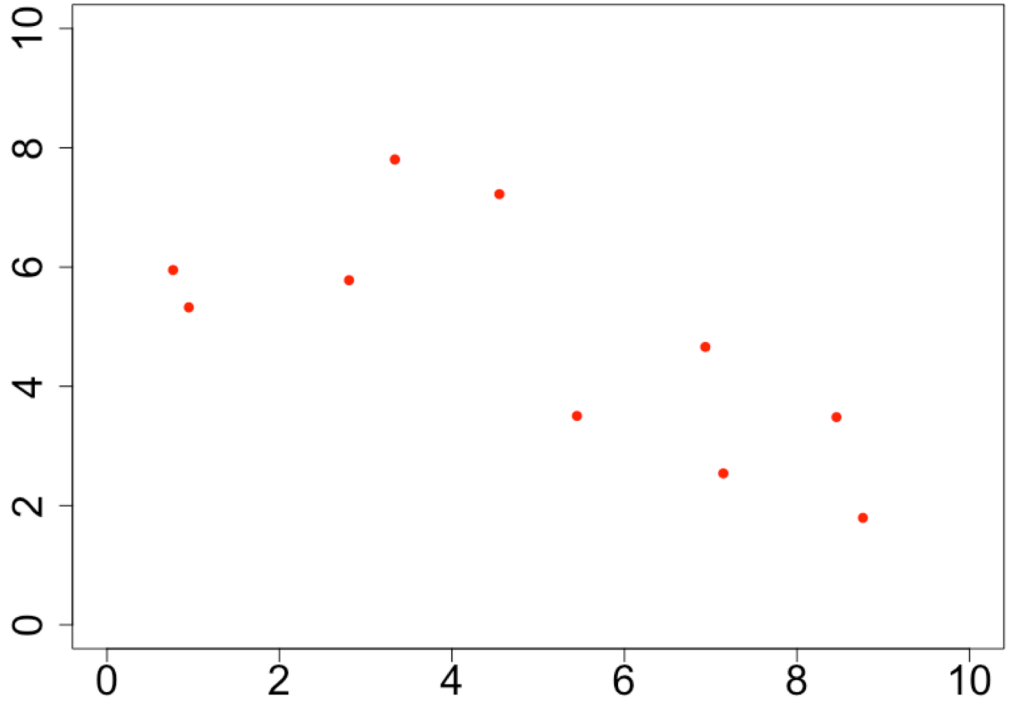
P(0) = 903/1000 = 0.903

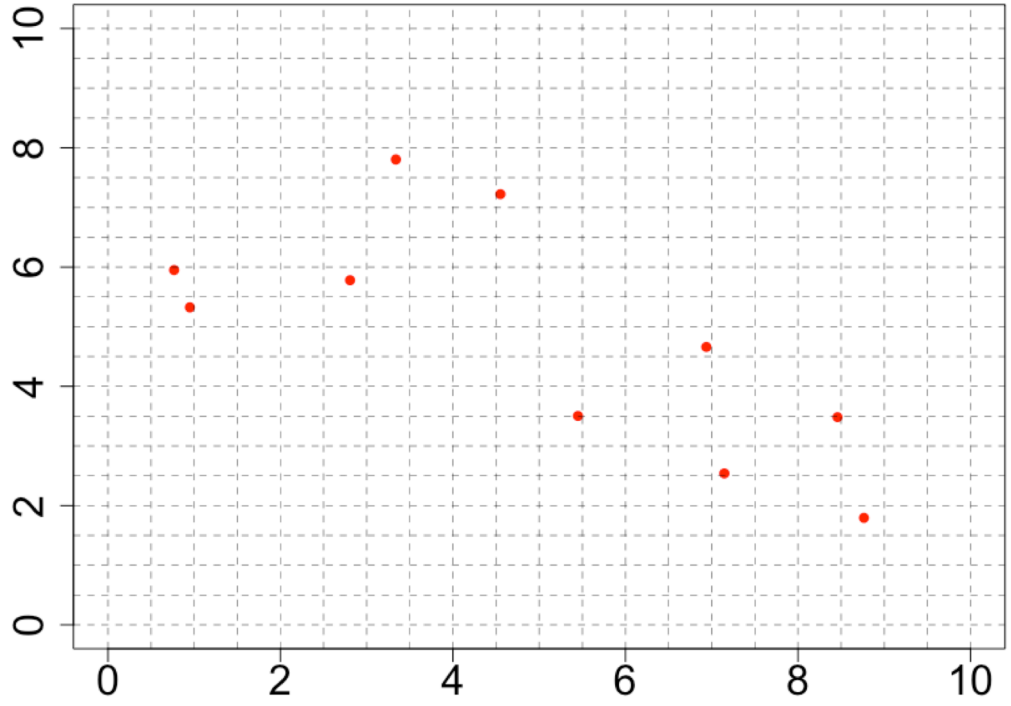
P(1) = 94/1000 = 0.094

P(2) = 3/1000 = 0.003

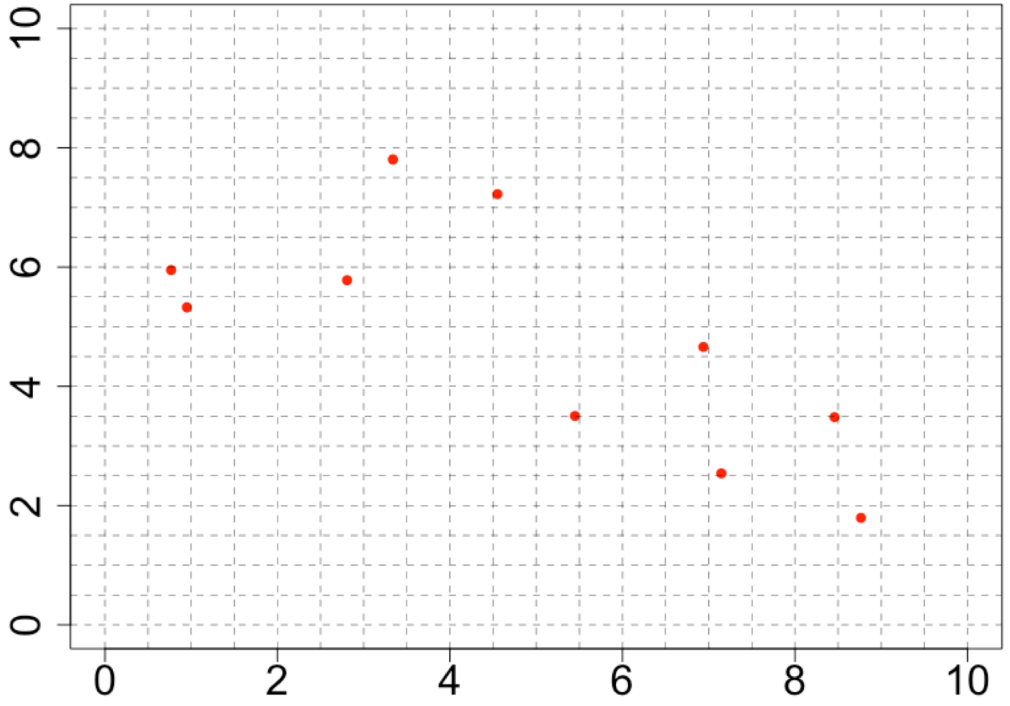
# From simulation of events to a probabilty model

# Zooming in : a probability model for random events in microscopic space units



# Zooming in : a model for random events in microscopic space units




We take the area and **we slice it N micro units** (N BIG)

**Focus on 1 micro unit**

In every micro unit **things happen rarely**

P(0 event) = 1- p

P(1 event) = p

P(> 1) = 0 (it's so rare)

X records how many things happened in the N units

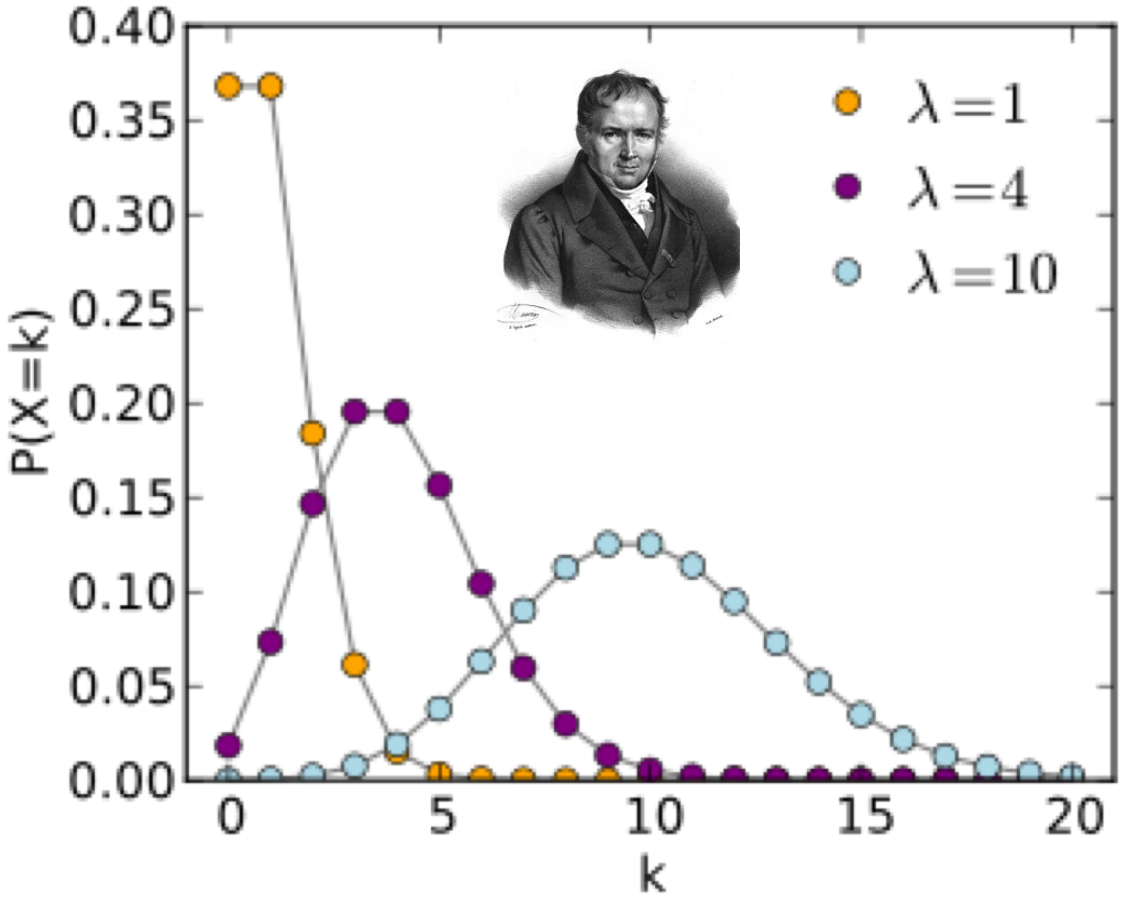
**X is BINOMIAL N trials and probability p**

E(X) =Np V(X) = N p (1-p)

If p becomes small and N larger (N*p = $\lambda$)

→ Then X is ….**POISSSON ($\lambda$)**

# The Poisson distribution




**X is the number of events** in a time/space interval

Events are happening at a **constant** rate per unit

Events are **independent** and their occurence do no affect each other

X is a **random variable**

X in {0,1,2,…}

P(x events)= $\exp(-\lambda)\ \lambda^x / (x!)$

**E(X) = λ**

**V(X) = λ**

λ measures how many events you expect ”on average”

λ > 0 but can be <<<1 or > 1

# How do we use the Poisson distribution in practice

**Calculate the probability of specific events**

We sequence a genome with coverage ”3 X”, we have on average sequenced every position 3 times.

What fraction have we missed?

The number of sequences ”landing on a position” **is Poisson ($\lambda$=3)**

P(0 event) = exp(-$\lambda$) = exp(-3)

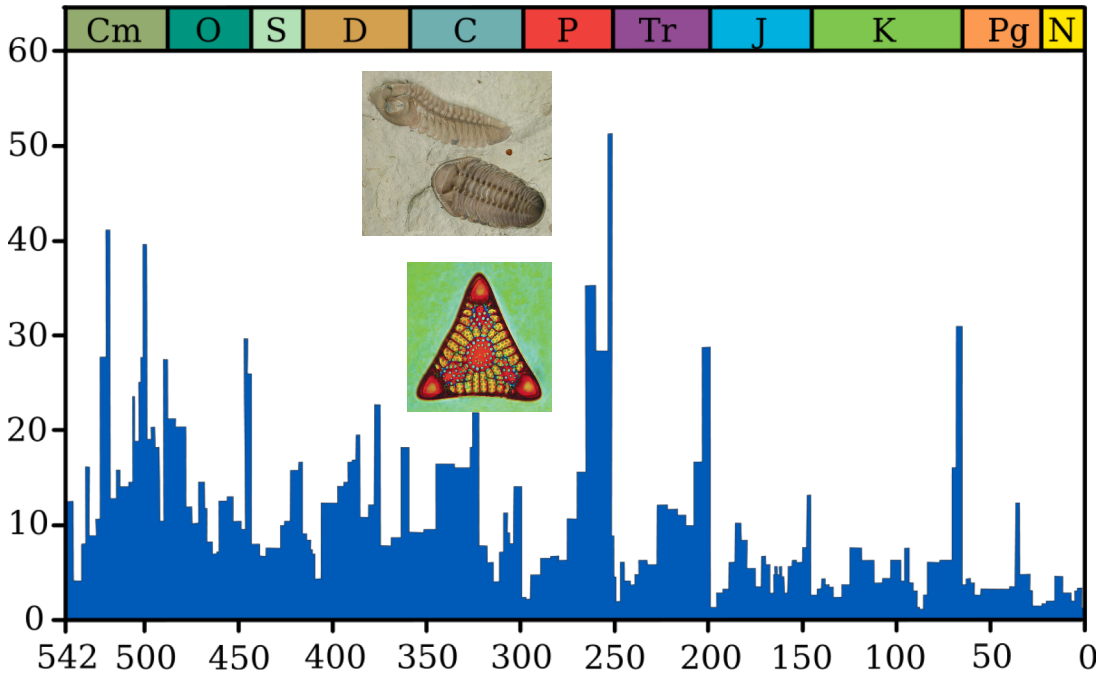
In R

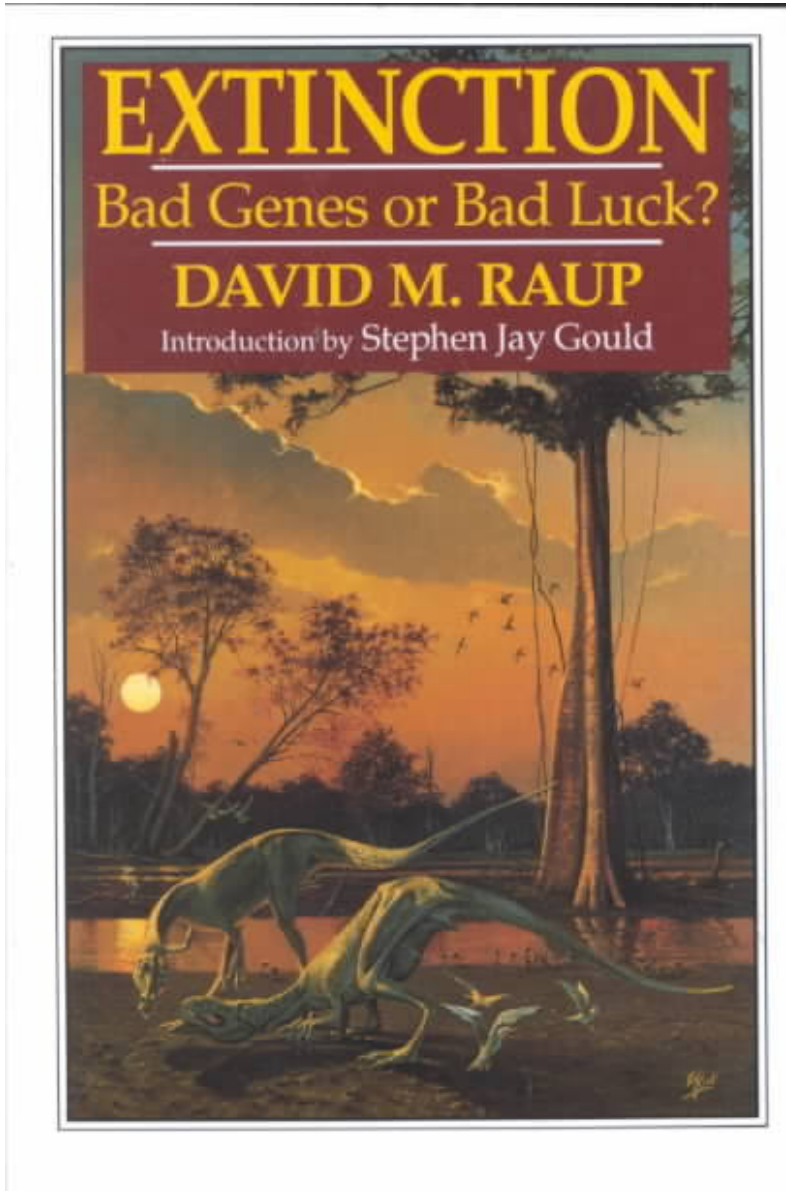
dpois(**x = 0**, lambda = 3) = 0.049787

**Use as a null model for randomly distributed events**

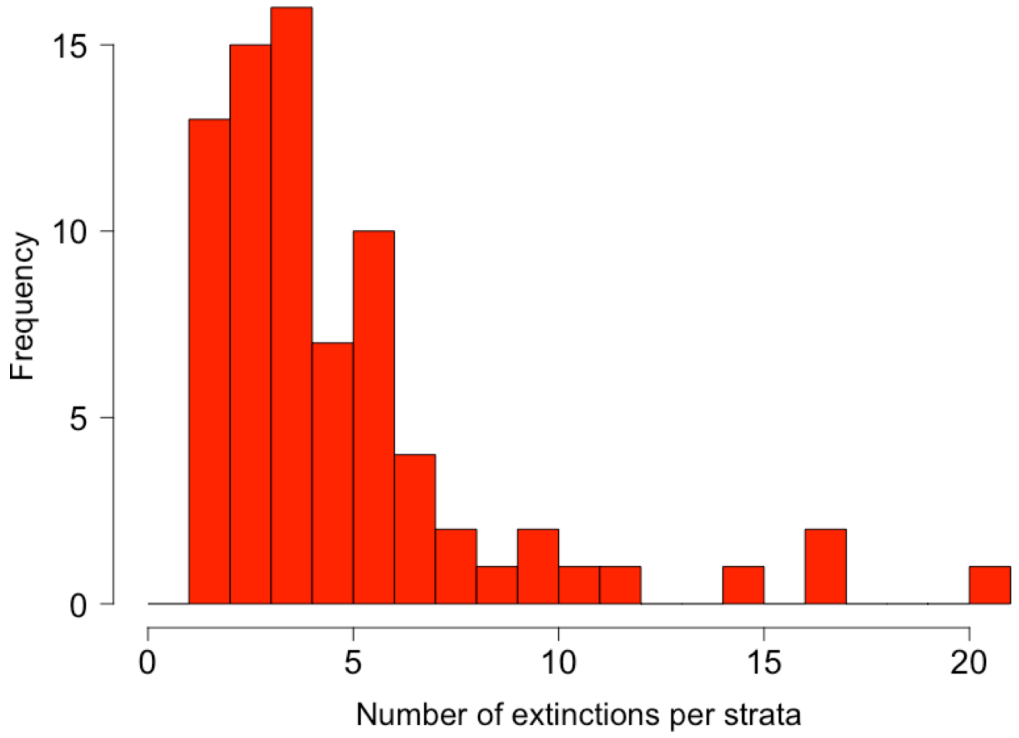
The Extinction data (Example 8.6)

# Extinctions in the fossil record : *Bad Genes or Bad Luck ?*



# The extinction data: observed distribution




76 strata

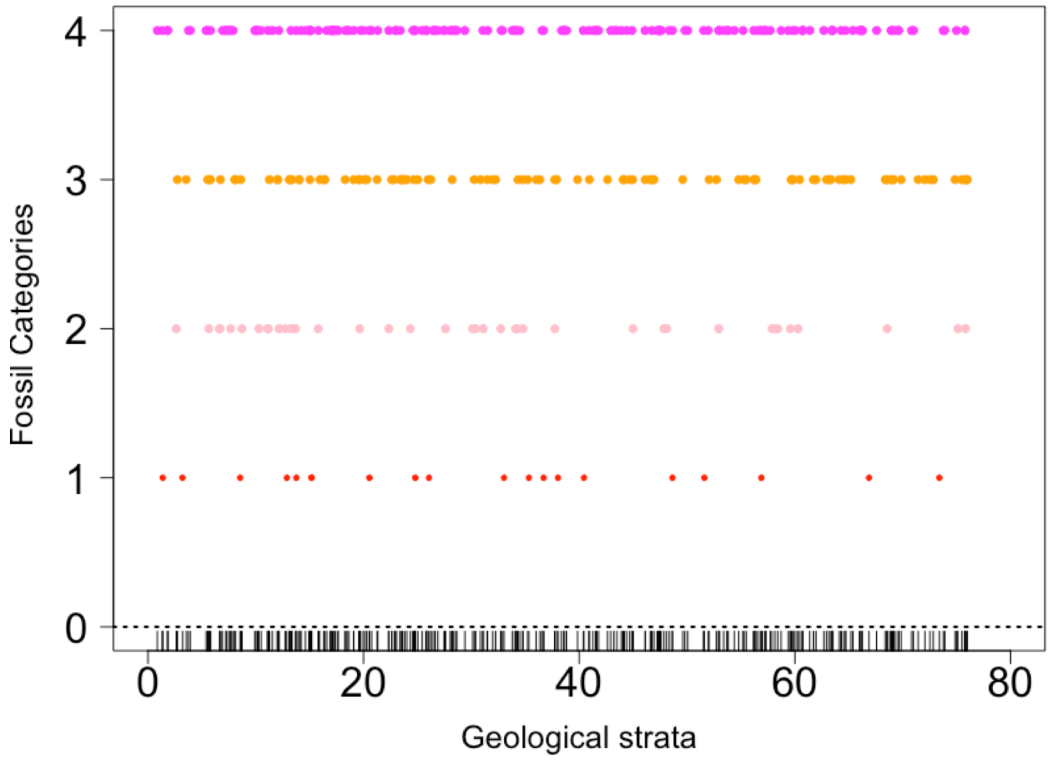
Total number of extinction events 320 (summed over all strata)

Mean per strata = 4.21

What should it look like

”at random” ??

# Lets vizualize the extinctions along the strata




For vizualizing

4 categories of fossils

Common,

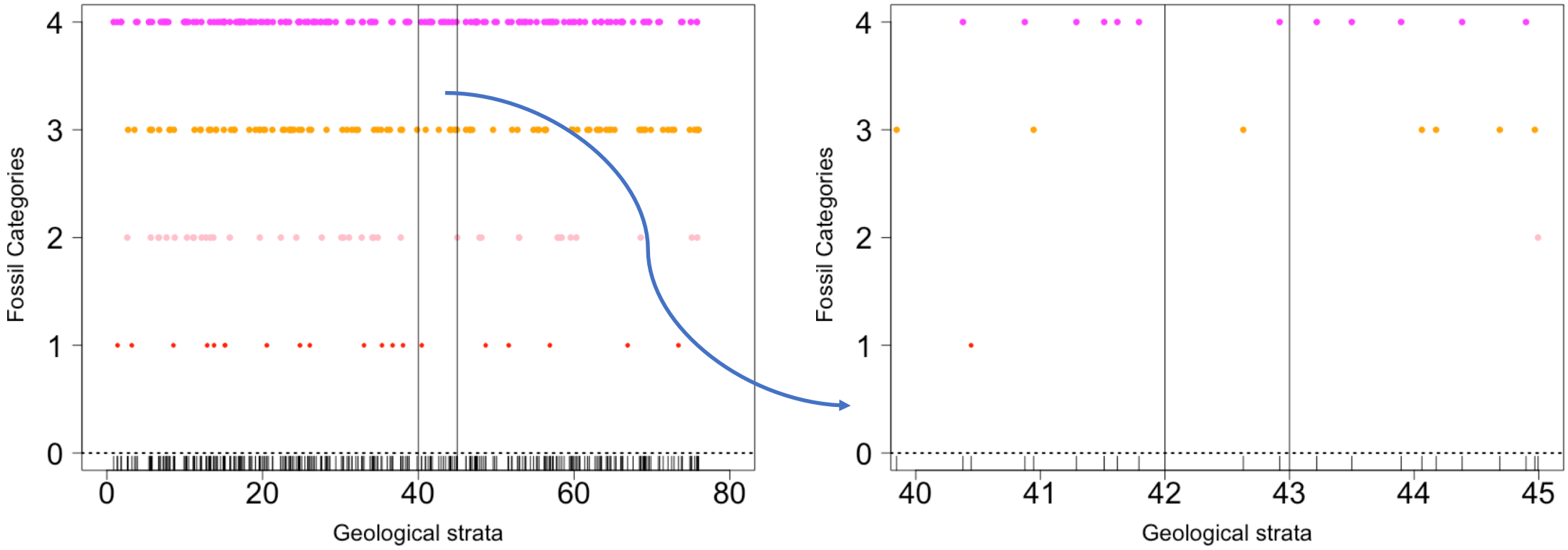
Uncommon,

Rare

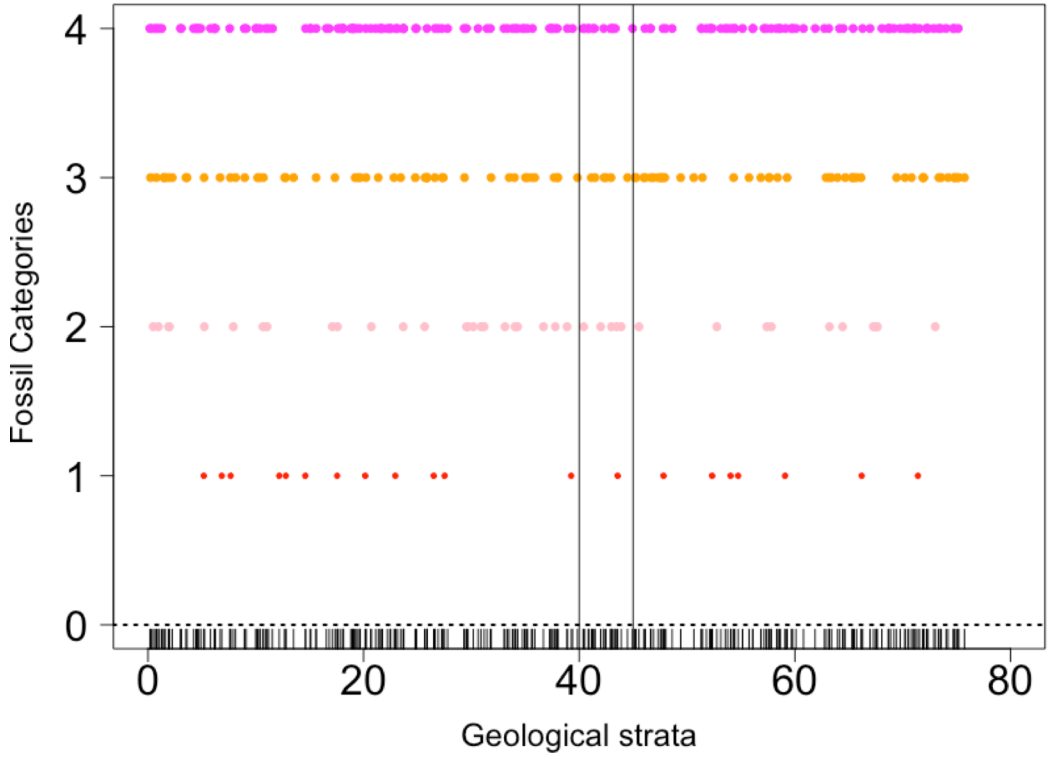
very rare

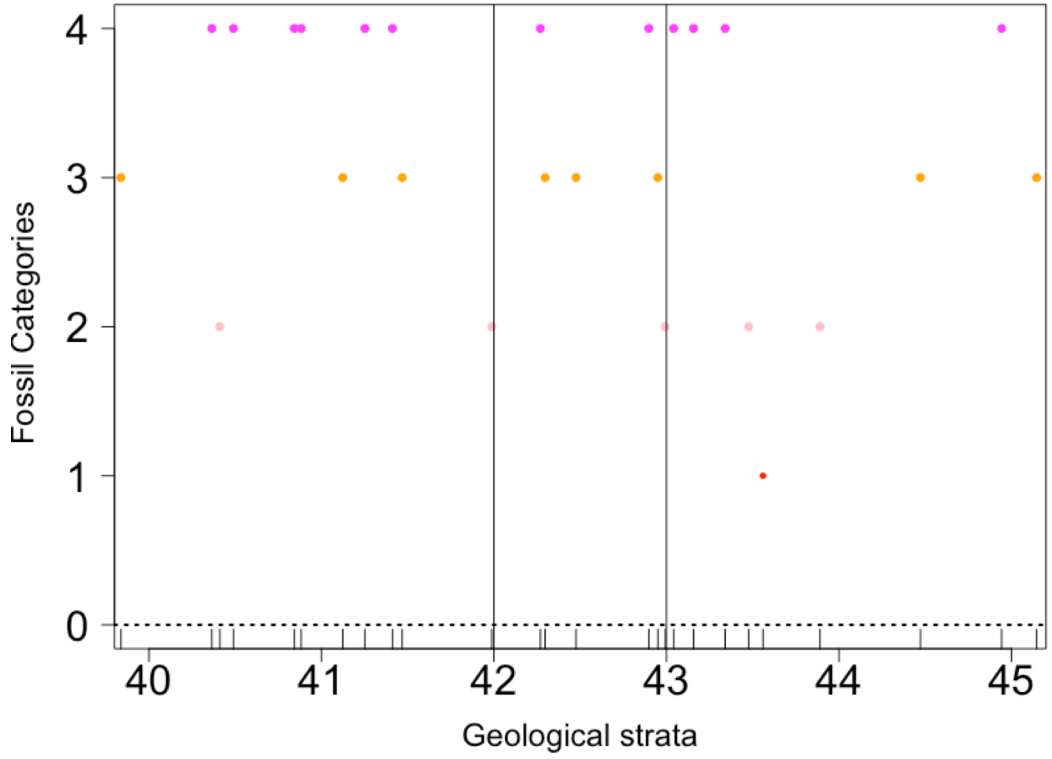
**In total 320 events of extinctions… to ”distribute” in 76 strata**

# Playing the tape of evolution, and zoom

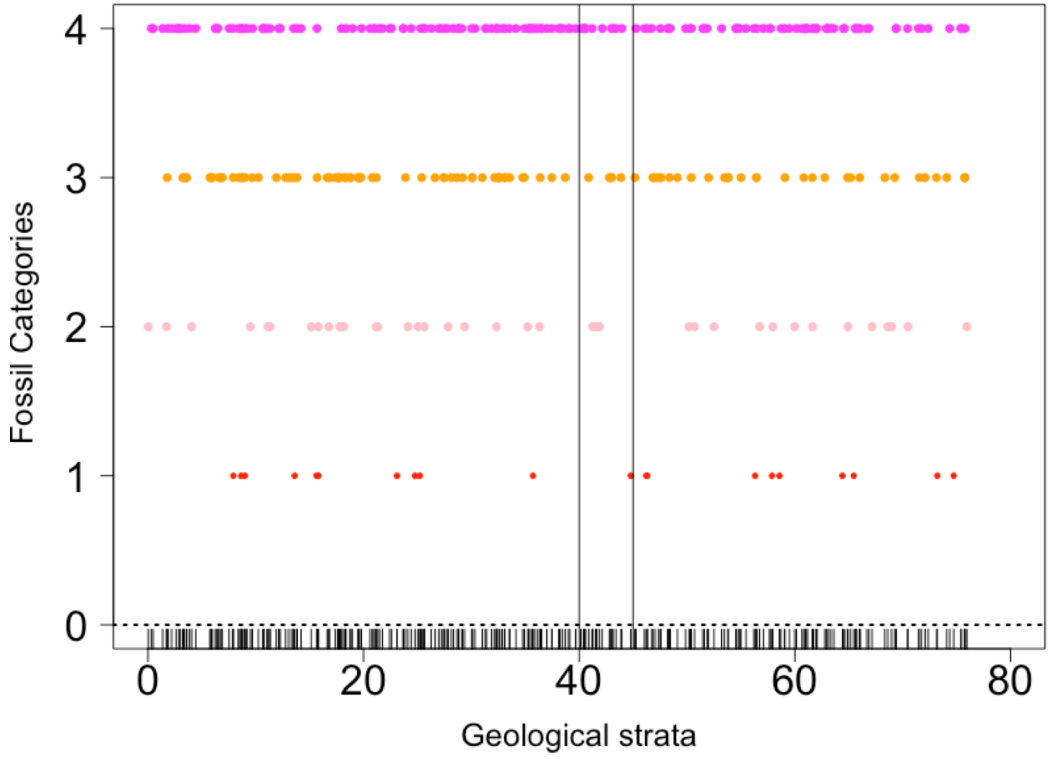


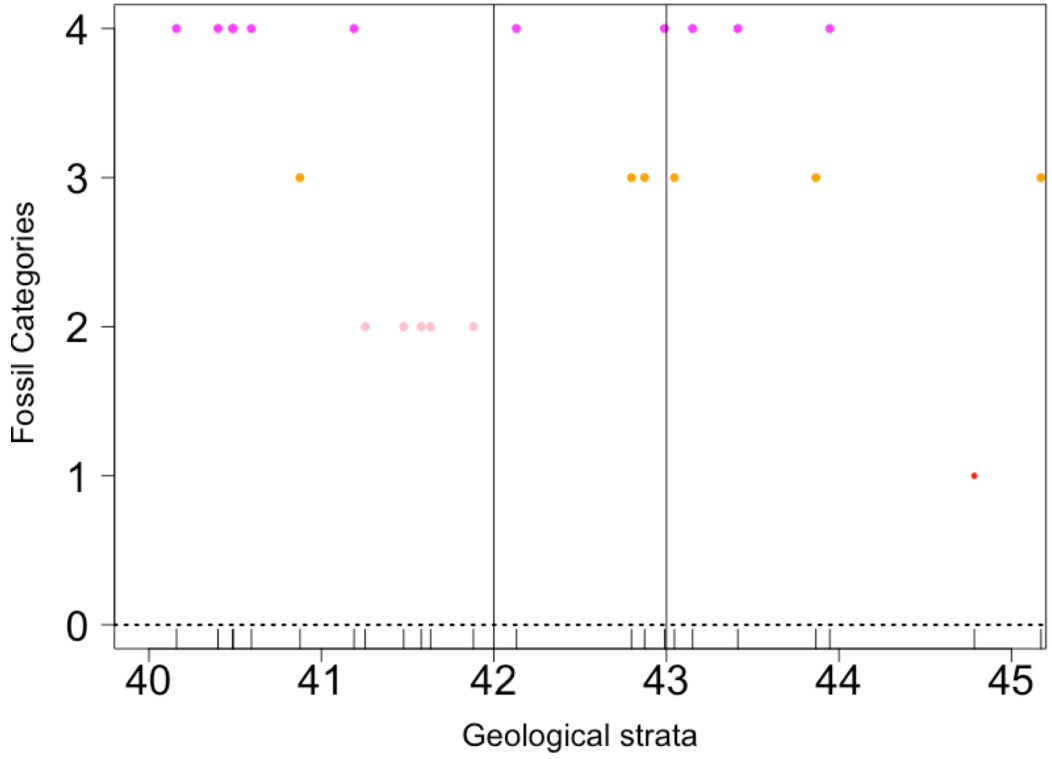
# Let’s replay ”the tape of evolution”



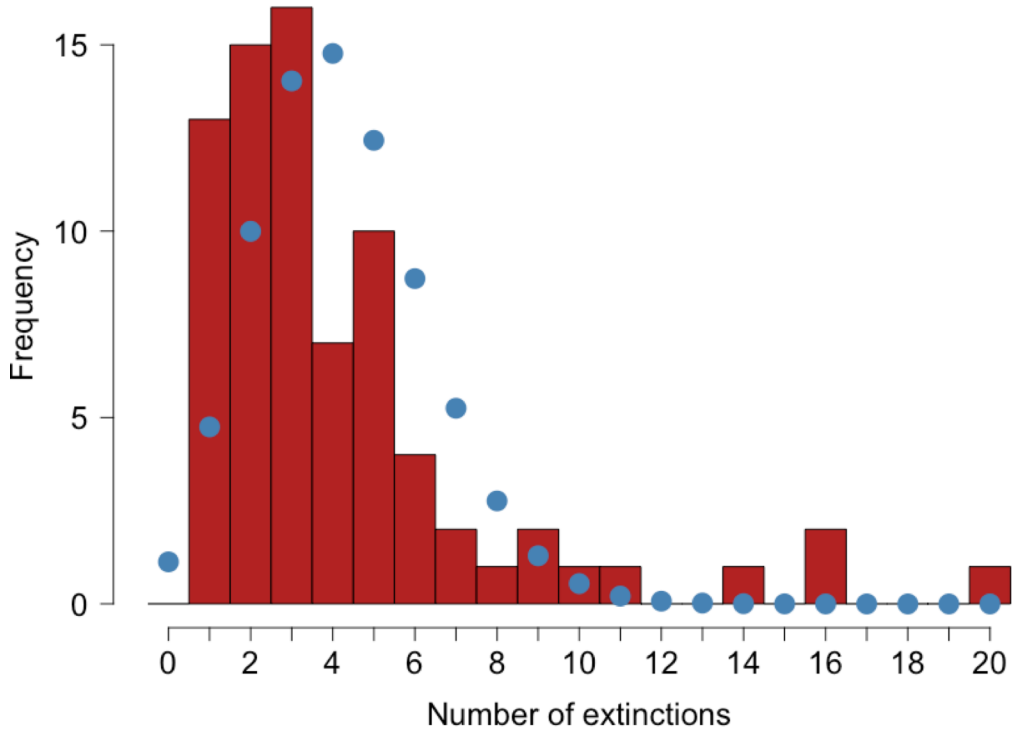



# Let’s replay ”the tape of evolution” … again …



# Using a GOF test for the data (1/3)




H0: ”data is random”

H0: ”Data is Poisson distributed”

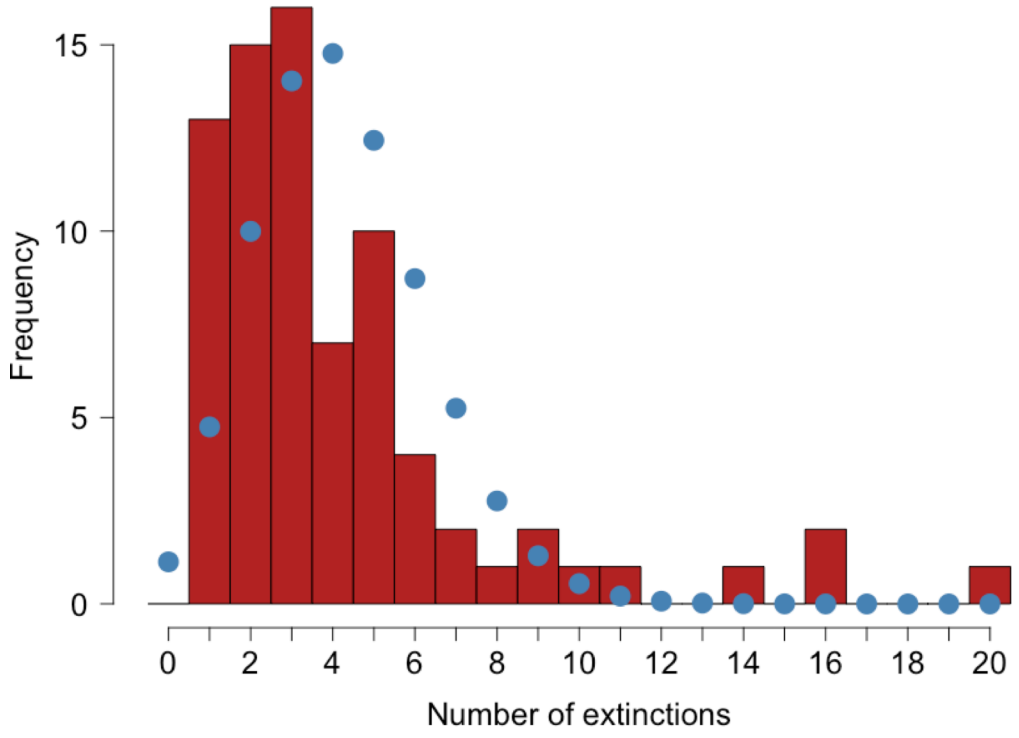
Choose the Poisson distribution that can plausibly account for the data ($\lambda = 4.21$)

Probability that there is 0,1,2, …20 extinctions per strata under H0

Expi = P( i extinction) *76
Expectation of counts for **”i”**

= P( i extinctions) *76

# Using a GOF test for the data (2/3)




H0: ”Data is Poisson distributed”

Poisson ($\lambda = 4.21$)

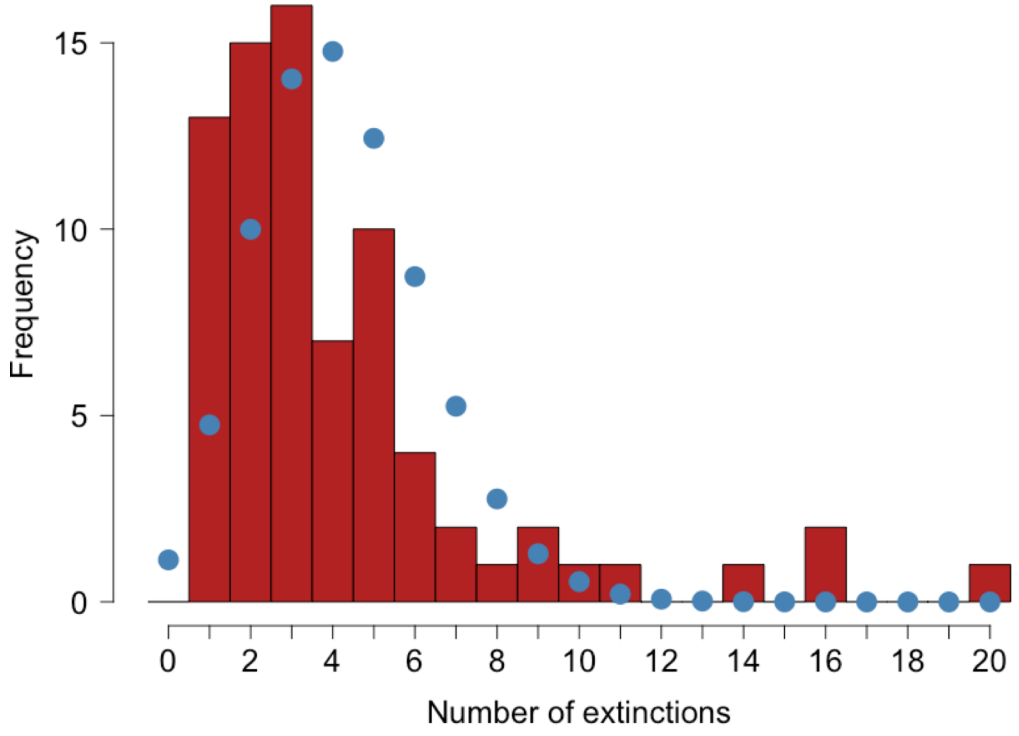
Probability that there is 0,1,2, …20 extinctions per strata under H0

Expi = P( i extinction) *76 Expi = P( i extinction) *76

How many classes ?

0 & 1, 2, 3, 4, 5, 6, 7, >7

# Using a GOF test for the data (3/3)




H0: ”Data is Poisson distributed”

Poisson ($\lambda = 4.21$)

Probability that there is 0,1,2, …20 extinctions per strata under H0

Expi = P( i extinction) *76

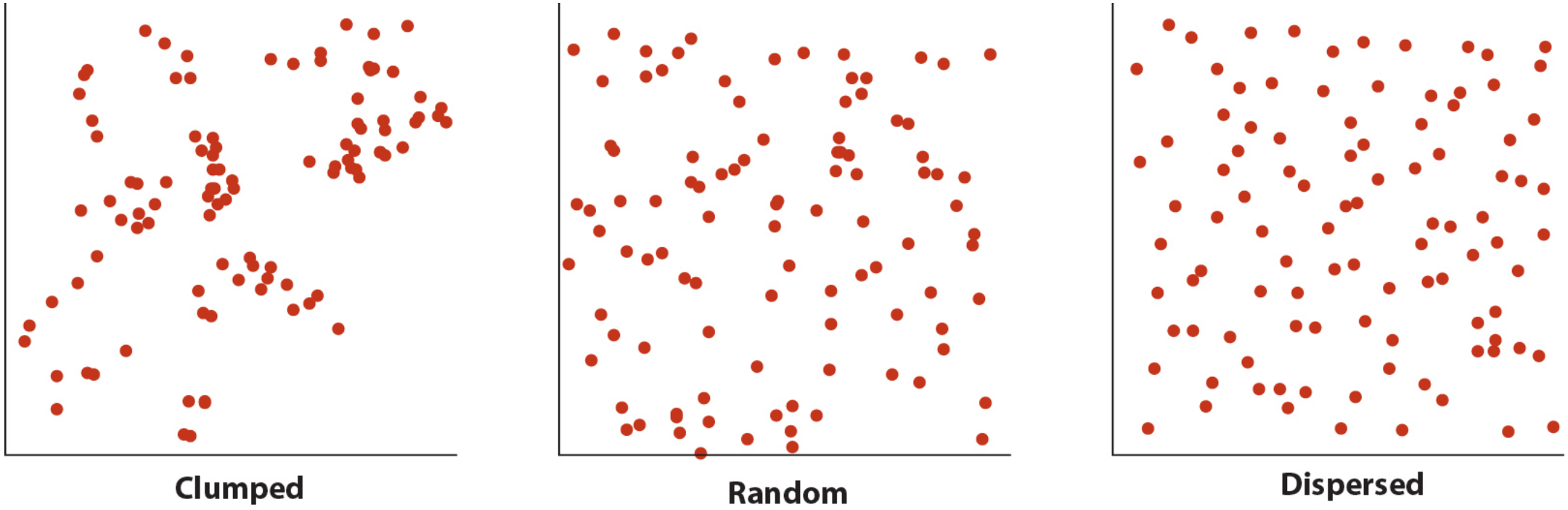
How many classes ?

0 & 1, 2, 3, 4, 5, 6, 7, >7

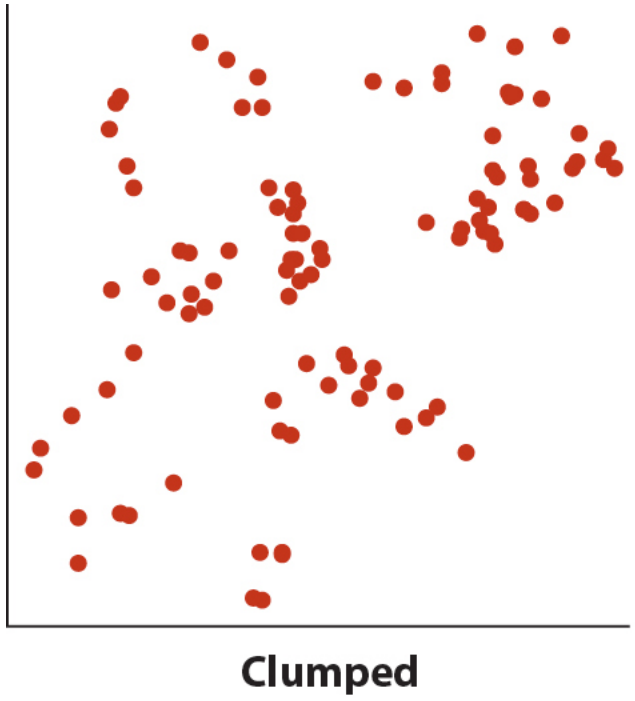
$\chi^2_{obs}$= 23.93 (8 classes)

d.f. = 8 -1 - 1

# Random in space/ time → Poisson distributed
# Poisson → Variance = Mean

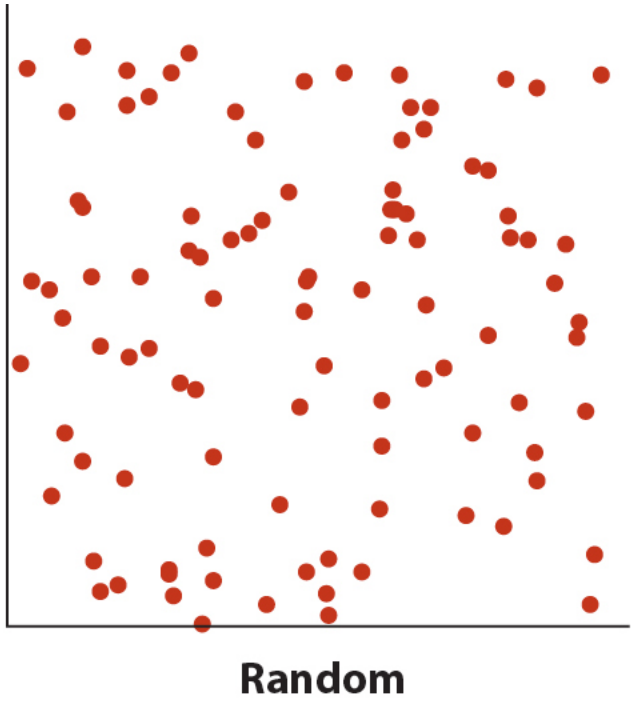


V/M >1



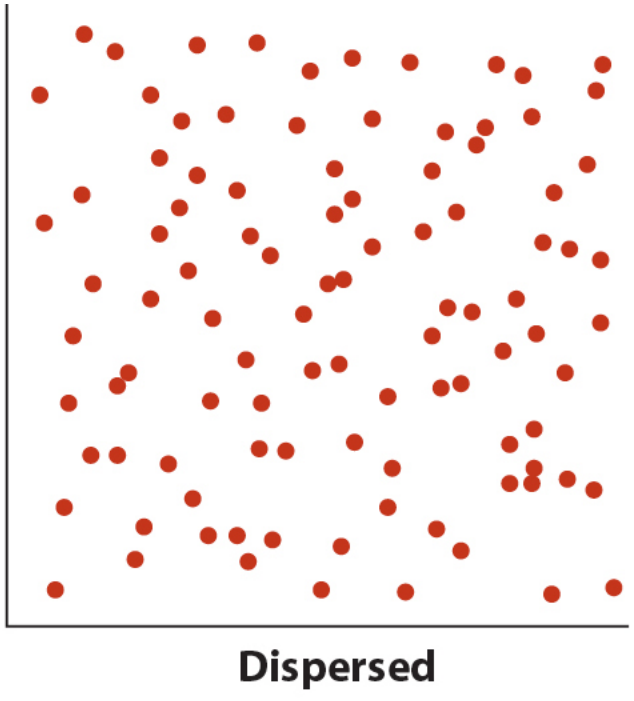
Clumped

V/M=1



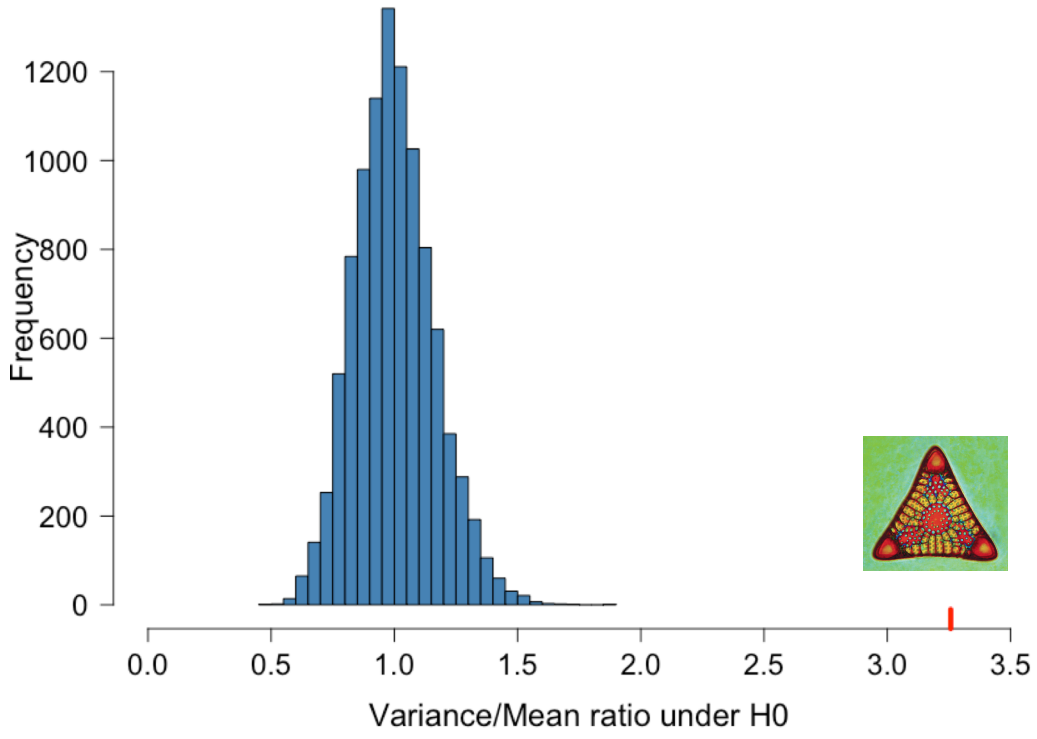
Random

V/M <1



Dispersed

# Using the Variance/Mean ratio to test if extinction data is random (poisson)




**M**ean in sample= 4.210526

**V**ariance in sample = 13.71509

V/M~3.26

10^6 simulations of data with n=76 observations

We simulate H0 for V/M:

- 76 independent observations Poisson distributed, with the fitted mean (4.210526)
- Calculate V/M each time

# RECAP 1 : ”if it happens randomly” then it is Poisson distributed

**X is the number of events** in a time/space interval

Events are happening at a **constant** rate per unit

Events are **independent** and their occurence do no affect each other

X is a **random variable**

X in {0,1,2,…}

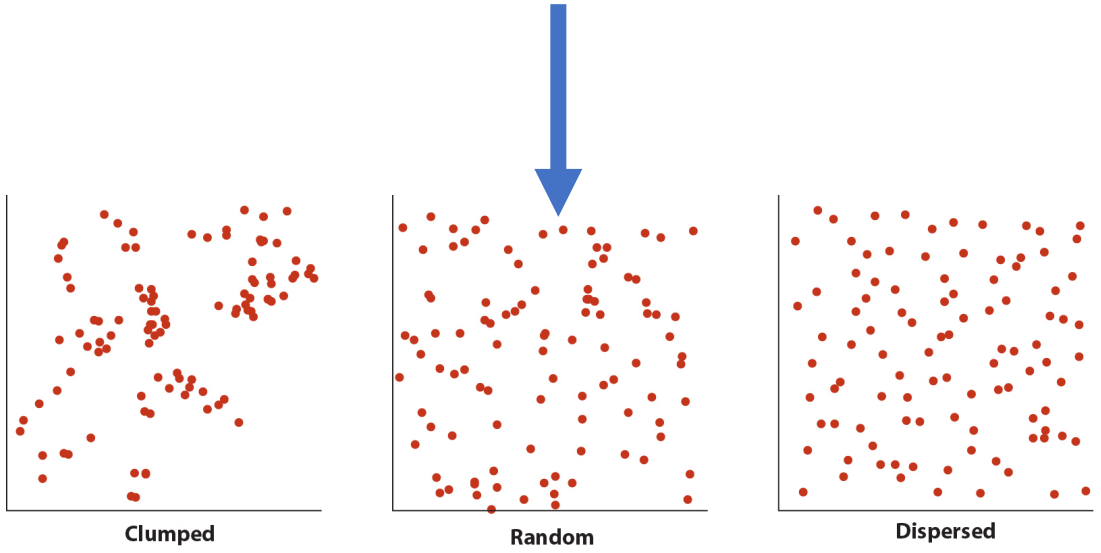
P(x events)= $\exp(-\lambda)\ \lambda^x / (x!)$

**E(X) = λ V(X) = λ**

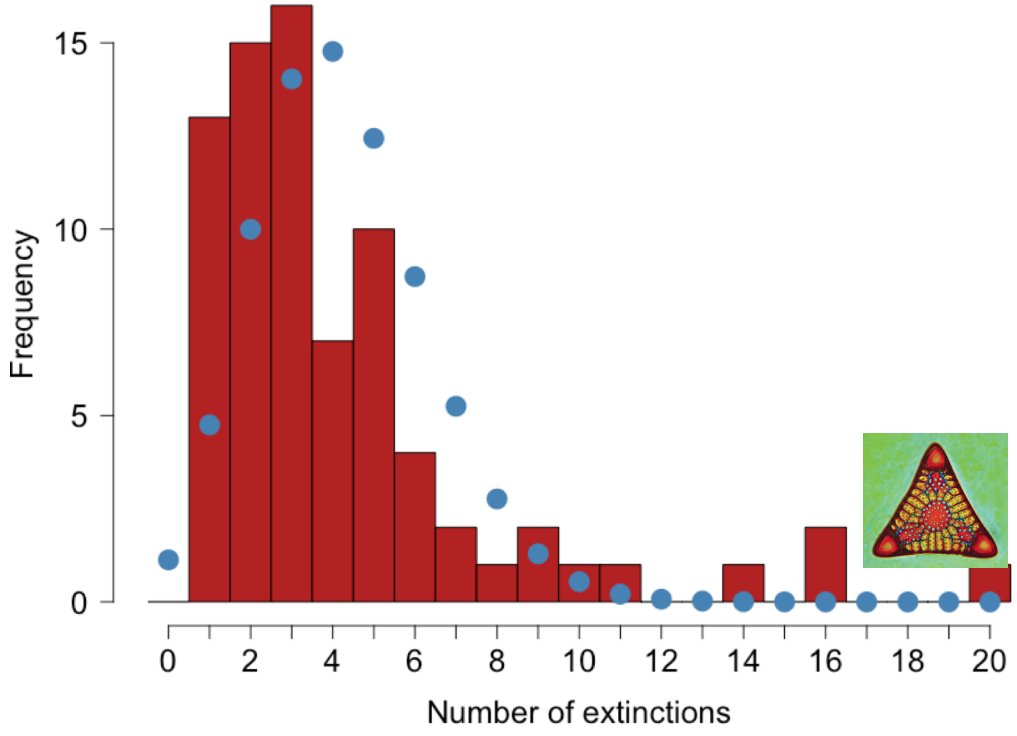
**M/V =1**

λ measures how many events you expect ”on average”

λ > 0 but can be <<<1 or > 1



# Recap 2
# Fitting data to a null model: poisson




A poisson distribution (**specified a posteriori**)

Choose the best poisson distribution (**fit** $\lambda$) and calculate expected counts using this poisson

Calculate the lack of fit ($\chi^2_{obs}$ )

(POOL if needed here 8 categories)

Compare with the null distribution for $\chi^2_{obs}$

*df* = **number of categories (8) - 1 -**

**df =6**

# Recap 3: Poisson (Random) or not?

RANDOM (POISSON)

Events  independent and homogenrous in rate

Counts of events are Poisson

Still need to know how rare are events (mean per unit)

Variance  = Mean  (V/M test)

Expected counts per category

DEVIATION FROM RANDOM

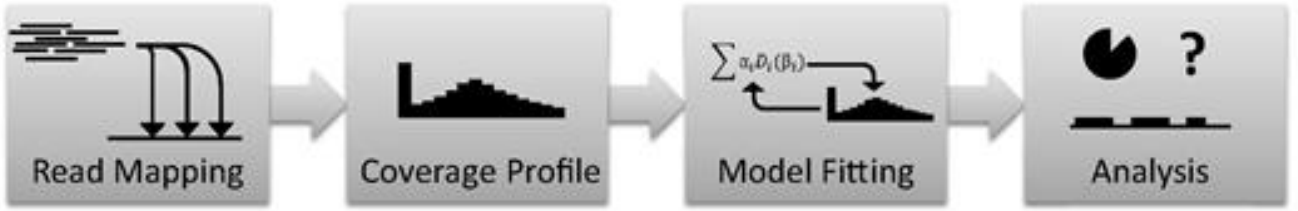
”Things become interesting”

→Heterogeneity in rates (what process is operating to drive it?)

→Non independence of events (what is causing it ?)

# Genome Coverage and Poisson distribution

Human genome is 3. $10^9$ nucleotides




Lets say you have done a volume of sequencing that ensures 5X coverage on a patient

Q1

P (missing Autosomal gene ) = p(0 coverage ) = ?

Q2:

What P( missing X linked gene ) ?

# Number of mutations in the genome

Human genome size is about

G=3 $10^{9}$ nucleotides

If Mutation rate per nucleotide per generation is m = 2. $10^{-8}$

Per genome & per generation, we can say that we expect to have X mutations

X = {0, 1, 2 …}

E(X) = Gm = 3 $10^{9}$ * 2. $10^{-8}$ =60

If mutations happen independently and at the same rate in individuals

X POISSON with mean $\lambda$=60

**Q: is the data poisson distributed?**

| | Father's Age (yrs) | Mother's Age (yrs) | Paternal | Maternal | Combined |
|---|---|---|---|---|---|
| Trio 1 | 22 | 19 | **39** | **9** | 48 |
| Trio 2 | 23 | 20 | **43** | **10** | 53 |
| Trio 3 | 25 | 22 | **51** | **11** | 62 |
| Trio 4 | 36 | 32 | **53** | **26** | 79 |
| Trio 5 | 40 | 39 | **91** | **15** | 106 |