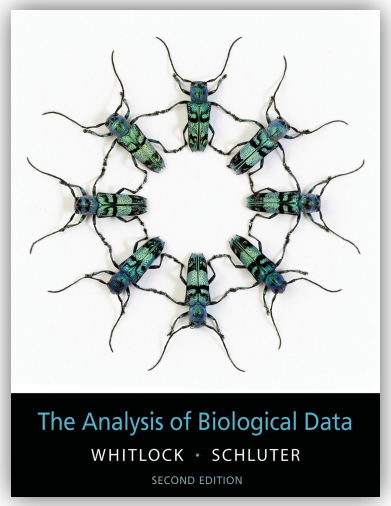
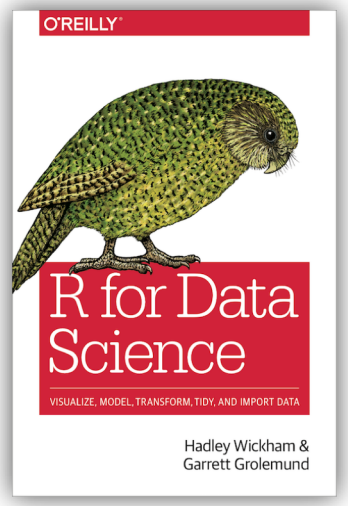
Data Science in Bioinformatics

week.04.

Palle Villesen & Thomas Bataillon





This week topics

- Hypothesis testing
  - H0 and HA
  - test statistic, distribution under H0 pvalue
  - What happens when we do several (many) tests
- SE and statistical test on proportions

- Back to the mammals dataset
  - more on X chromosomes
  - Summarizing data (again)
  - Binomial tests and their versatile use

The simplest example

- Wrap balloon around head of toad
- Which hand does it use to get if off?
- 18 toads
- 14 used right hand
- 4 used left hand
- What is H_0 ?
- What is H_a ?

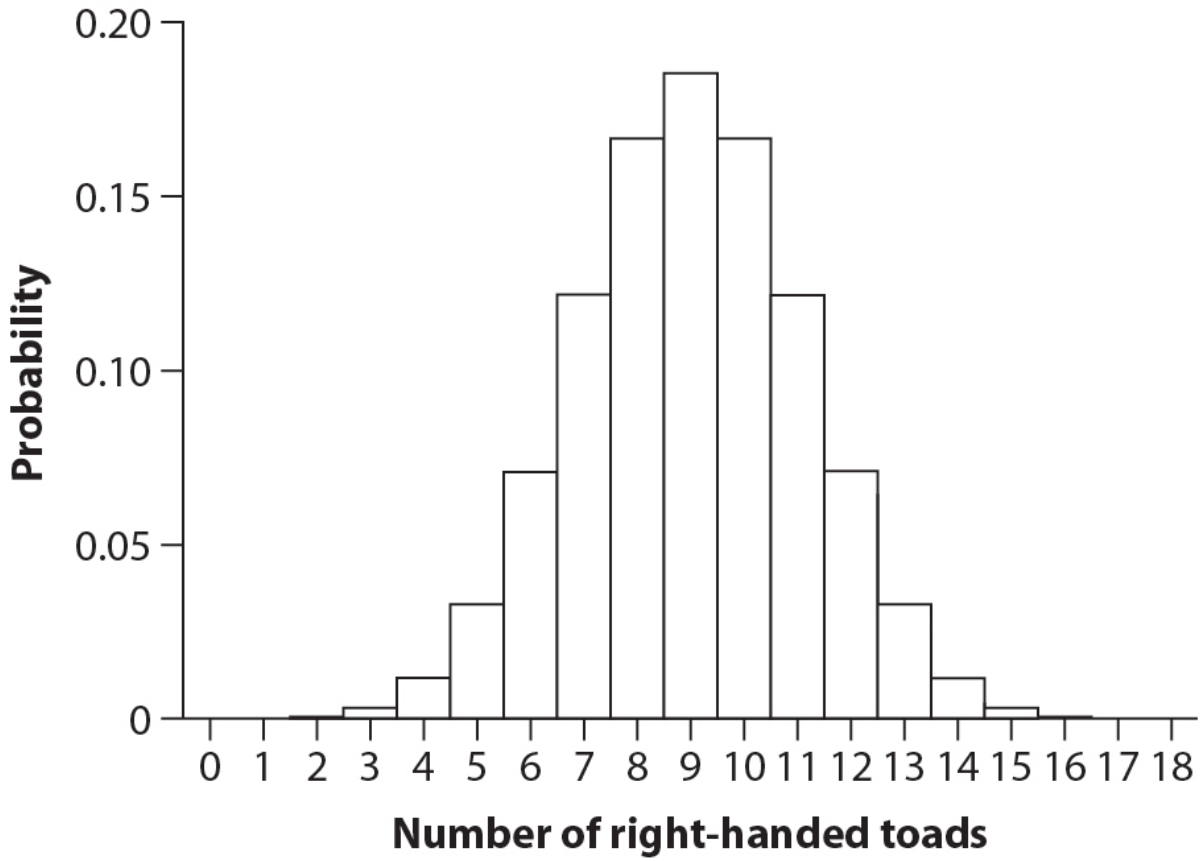
Test statistic

- A number calculated from the data we use to evaluate how compatible our data are with the expected result (assuming H_0 is true).
- So is 14 enough to say the H_0 is wrong?
- What is the expected test statistic in H_0 is true?
- What is the expected test statistic in H_0 is true and I tested 100 toads?

Null distribution

- If H_0 is true:
  - What is the probability of 0 right handed toads?
  - What is the probability of 1 right handed toads?
  - What is the probability of 17 right handed toads?
  - What is the probability of 18 right handed toads?

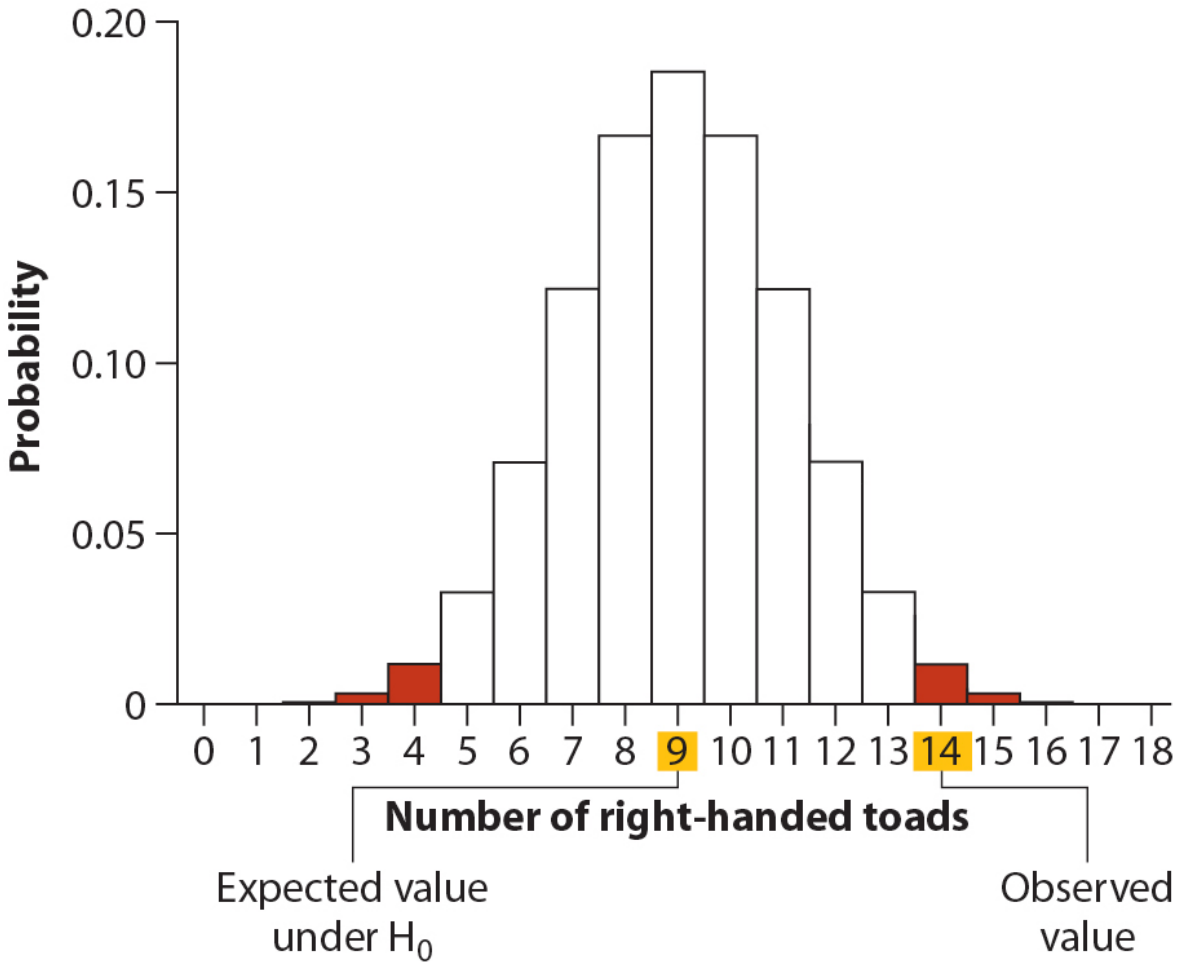
Null distribution for 18 toads



P-value

- The P value is defined as the probability, under the assumption of no effect or no difference (the null hypothesis), of obtaining a result equal to or more extreme than what was actually observed.
- p = probability of seing YOUR data assuming H0 is correct
- Low p value = surprise = H0 is probably NOT correct

P-value


0.20
0.15
0.10
0.05
0
Probability
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
Number of right-handed toads
Expected value under H_0
Observed value

P-value

Pr[14 or more] =

Pr[14] + Pr[15] + Pr[16] + Pr[17] + Pr[18]

Pr[4 or less] =

Pr[0] + Pr[1] + Pr[2] + Pr[3] + Pr[4]

Pvalue = Pr[4 or less] + Pr[14 or more]

~ 2 * Pr[14 or more]

= 0.031

A few questions that motivate the use of proportions

- Frequency of an allele at a SNP in a sample
- Proportion of individuals affected by a condition
- Many other examples where you can categorize data/ outcome in two categories:

  Low gene expression versus ”the rest”

  High dn/ds, etc.

The basic statistical questions about the proportion in a sample

Up to now, we focused a lot on MEANS

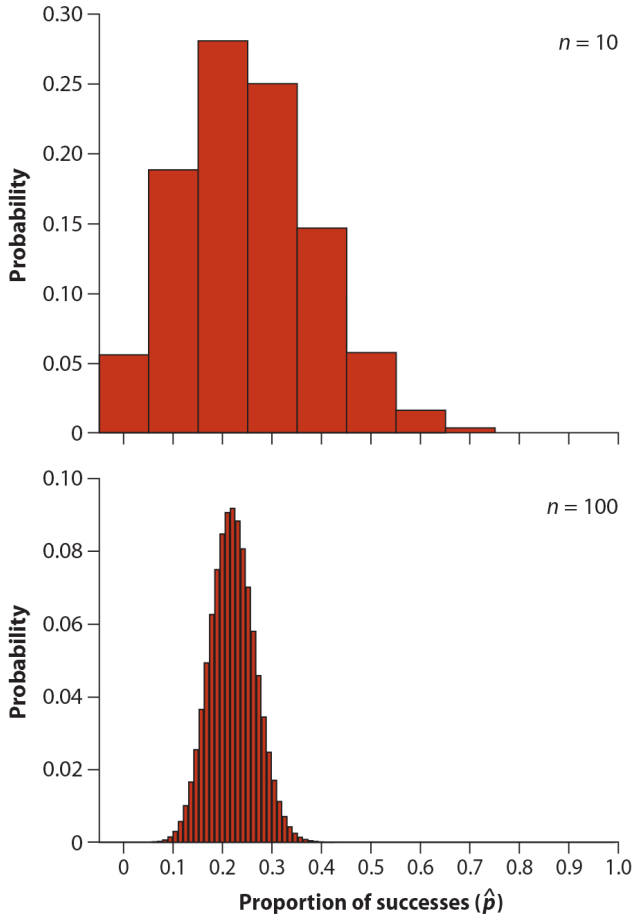
- Mean in a sample as sensible way to guess the mean in a population
- Its sampling distribution → SE
- A confidence interval

Now we are interested in a proportion

- Can we estimate the proportion in a population from a random sample ?
- If so how does the sampling variance ”behaves” ?
- Can we get a confidence interval around a proportion?

- YES WE CAN … because we can use the binomial distribution

The sampling distribution for a proportion



As with the mean in a sample

- Sampling variation is decreasing with n
- Can we say exactly how ?

From counts to proportions in a sample

Counts in a sample is a random variable

X is the number of occurence of a category in a sample of size **n**

Right versus Left

A versus T at a nucleotide, etc…

If the observations in the sample are independent and yield the chosen category with identical probability p

→X is binomial(**n**,p)

We know that

E(X)= **n**p

V(X)= **n**p(1-p)

Proportion in a sample is also a random variable

p^= X/n is a random variable recording the observed proportion in a sample

E(p^) = E(X) /n (1/n is a constant)

So **E(p^)=n p / n = p**

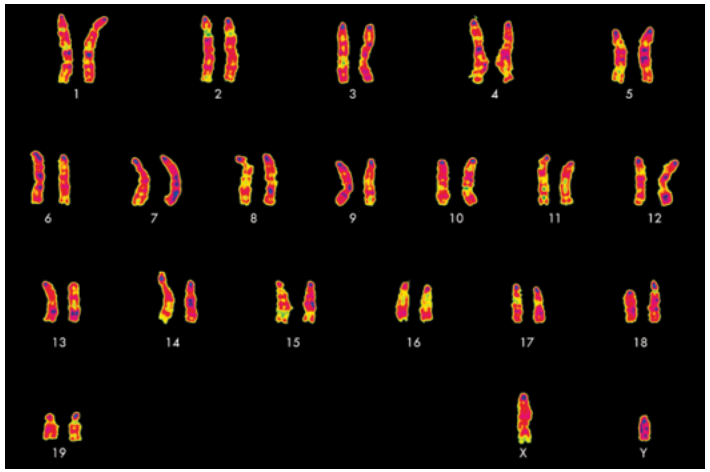
→ p^ is an unbiased estimator of p

V(p^) = V(X) / n^2 (1/n is a constant)

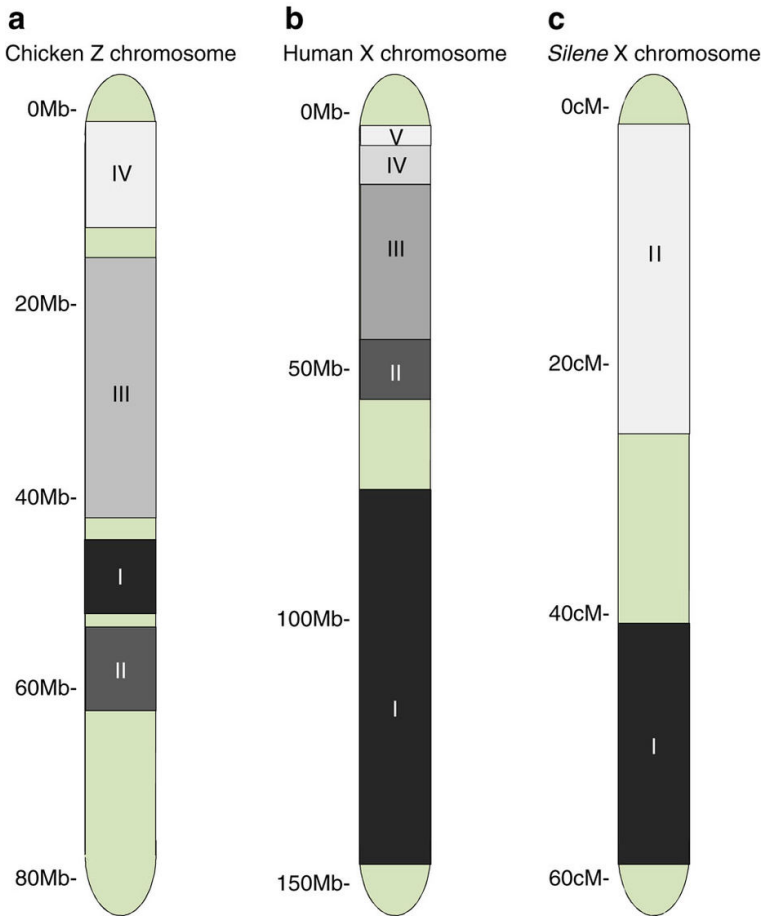
So **V(p^) = p(1-p)/n (See p185)**

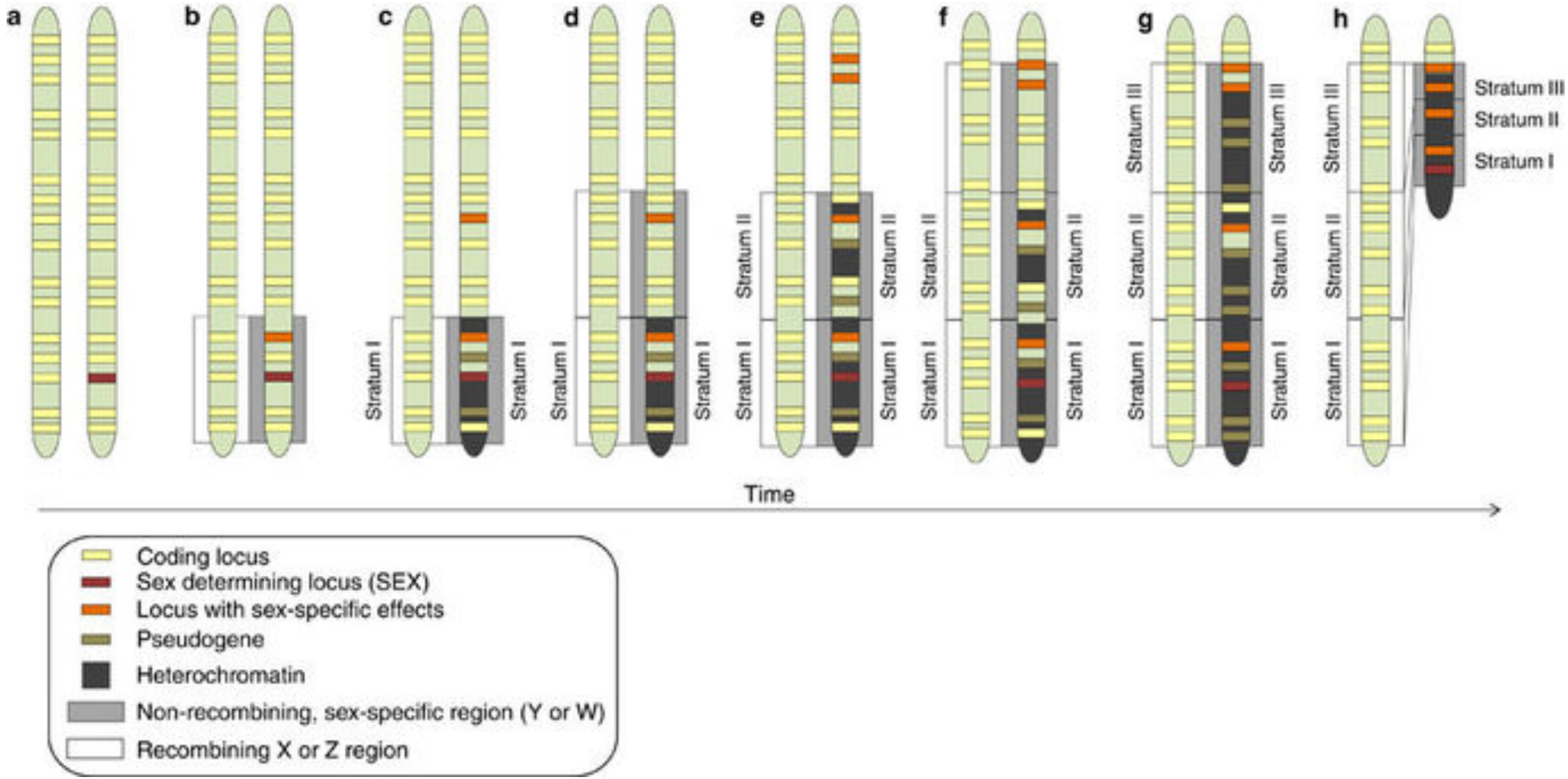
→ p^ has a sampling variance shrinking with ***n***

The mysterious X chromosome meets the binomial test

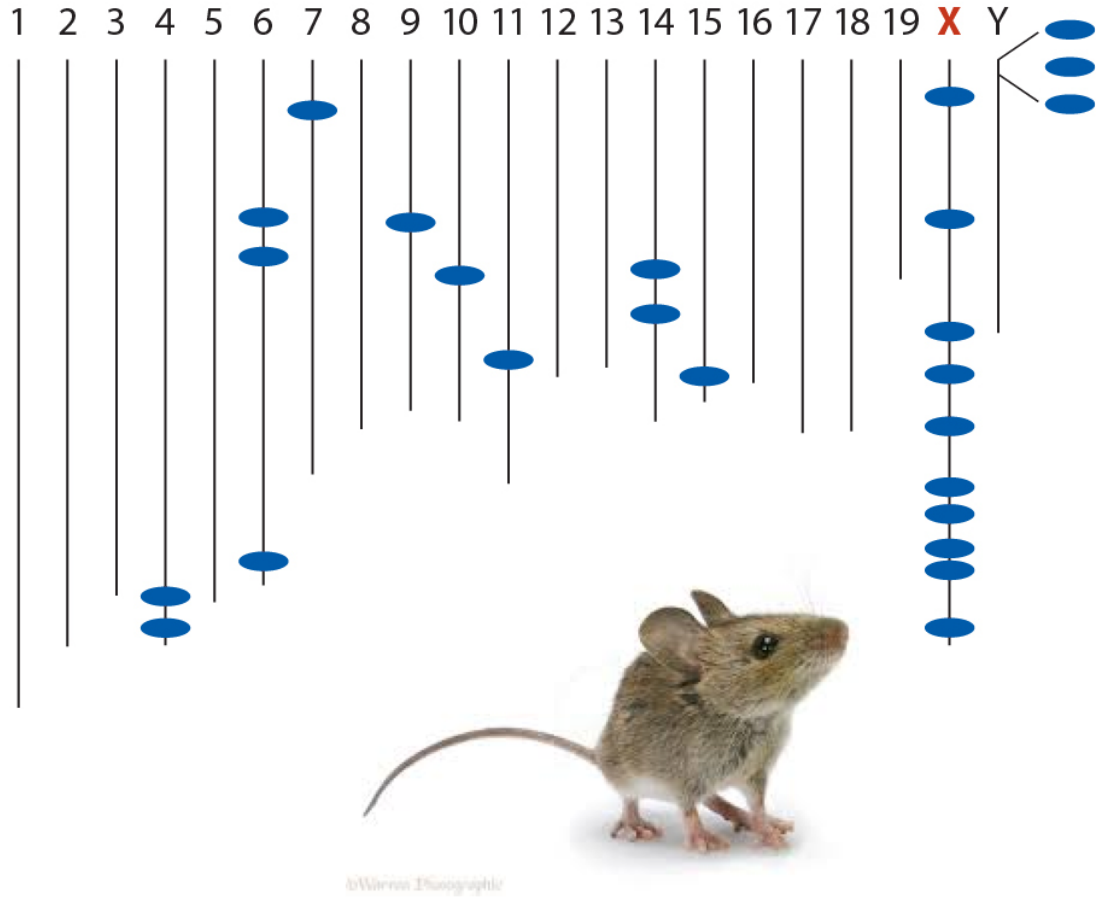


A few fun facts about X chromosomes and their evolution




a
b
c
d
e
f
g
h
Stratum I
Stratum II
Stratum III
Stratum III
Stratum II
Stratum I
Time
Coding locus
Sex determining locus (SEX)
Locus with sex-specific effects
Pseudogene
Heterochromatin
Non-recombining, sex-specific region (Y or W)
Recombining X or Z region

Is the X chromosome special ?



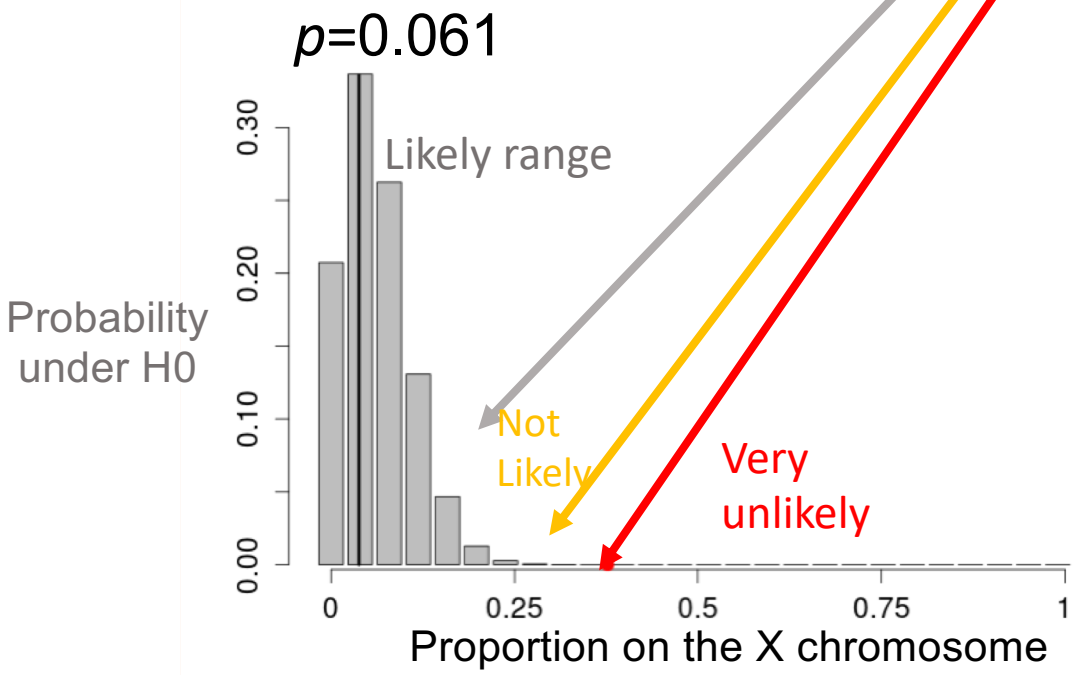
There are 25 spermatogenesis genes in the mouse genome and

10 genes are on the X chromosome

But, X chromosome only contains 6.1% of the total genes.

Null model (H_0): any gene has the same underlying probability to be on the X

Example: Are spermatogenesis genes disproportionately located on the X chromosome?



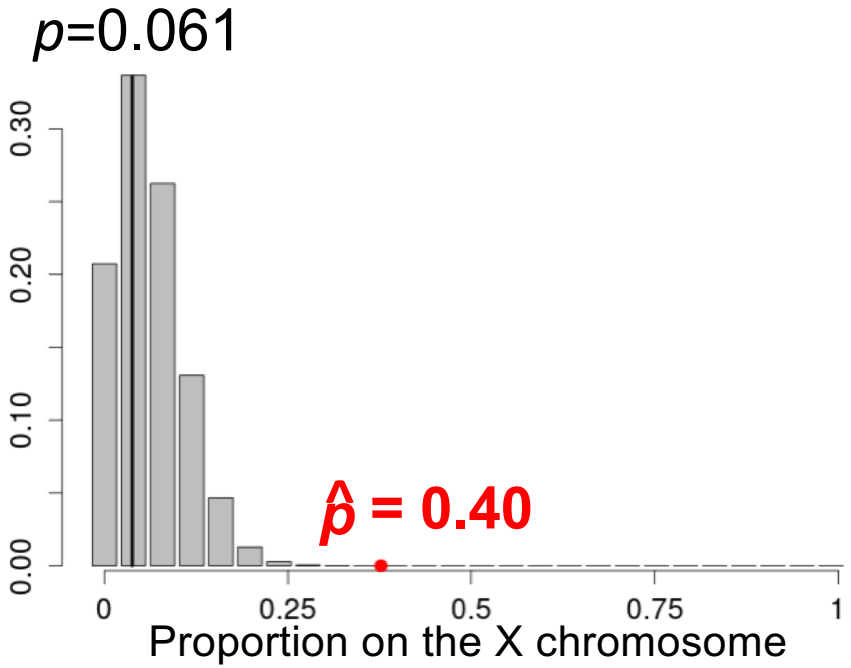
H_0: The probability that a spermatogenesis gene falls on the X chromosome is p=0.061.

H_A: The probability that a spermatogenesis gene falls on the X chromosome is something other than 0.061 ($p \neq 0.061$).

Example: Are spermatogenesis genes disproportionately located on the X chromosome?

Best estimate of proportion:

$\hat{p} = X/n = 10/25 = 0.40$



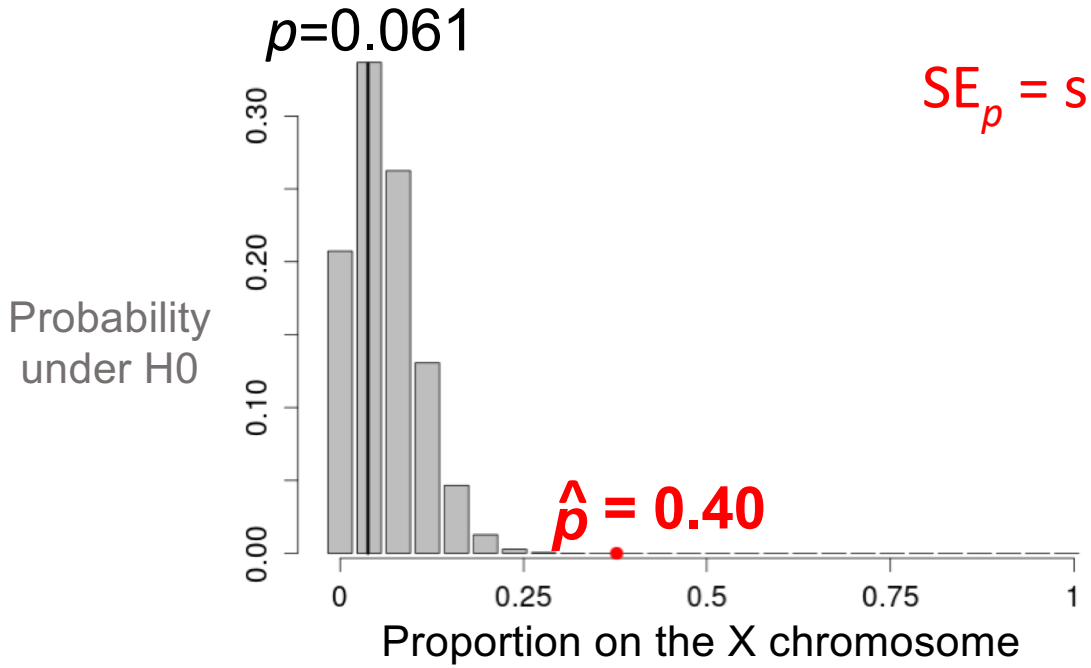
Example: Are spermatogenesis genes disproportionately located on the X chromosome?

Our estimate of proportion:

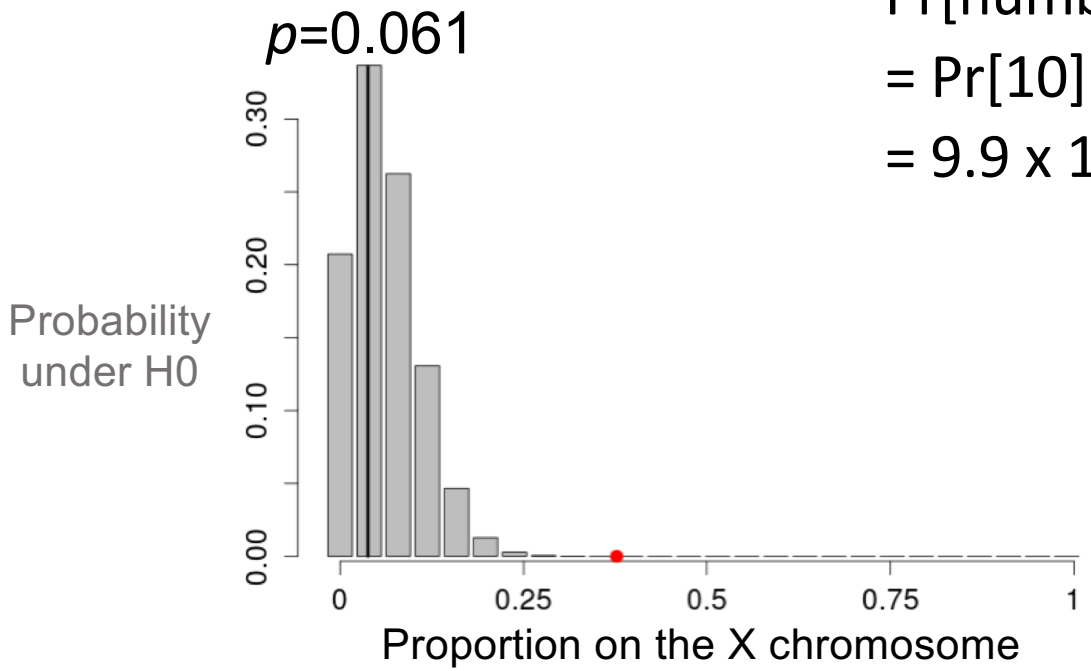
$\hat{p} = X/n = 10/25 = 0.40$

The standard error of proportion:

$SE_p = \text{sqrt}(0.4\ (1\text{-}0.4)/\ 25) = \ldots$



Example: Are spermatogenesis genes disproportionately located on the X chromosome?



Now find the P-value:

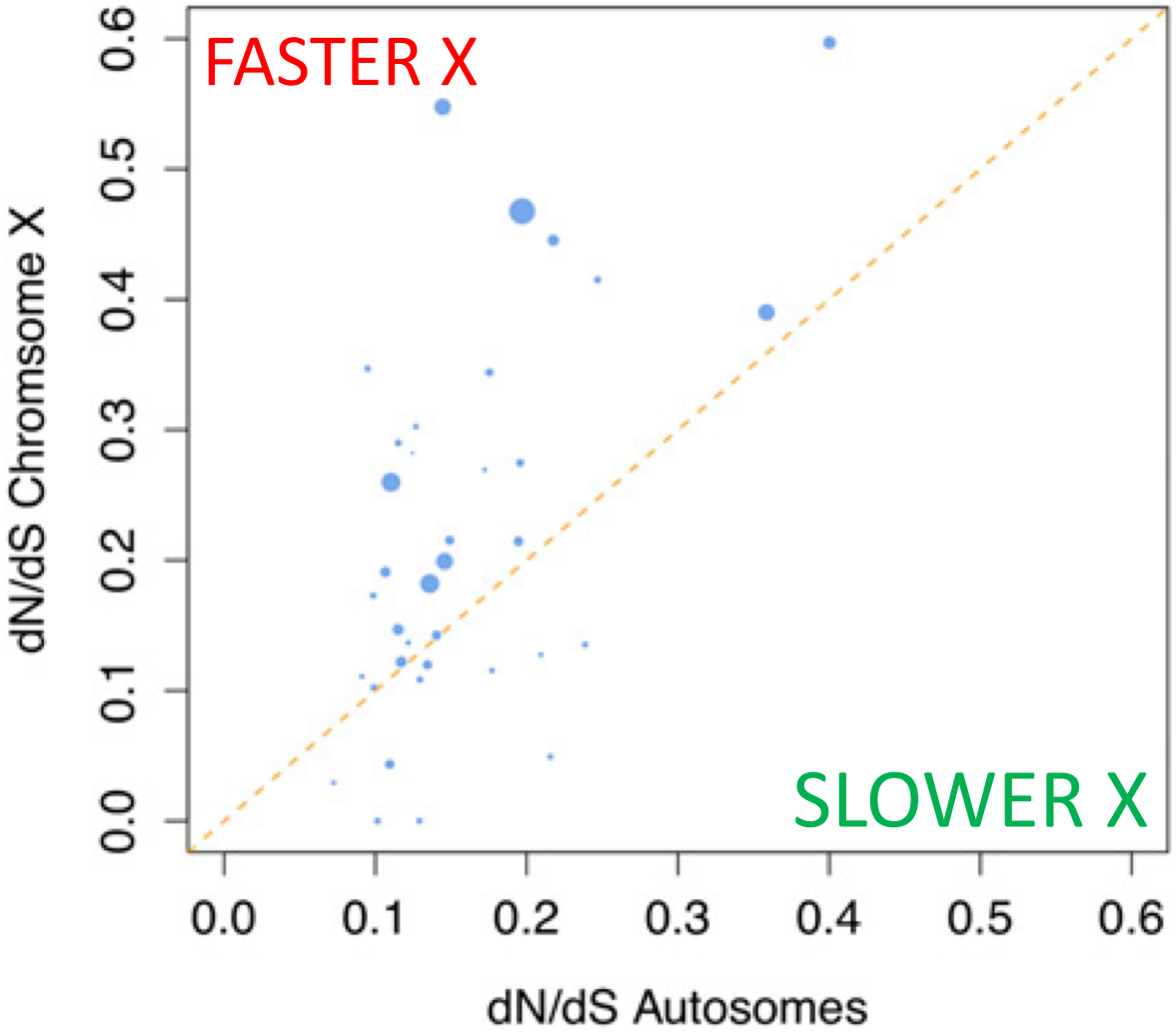
Probability of getting 10 genes or more on the X "by chance alone" (under H0)

Pr[number of successes ≥ 10]

= Pr[10] + Pr[11] + … + Pr[25]

= 9.9×10^{-7}

More about the X …A versatile use of the binomial test

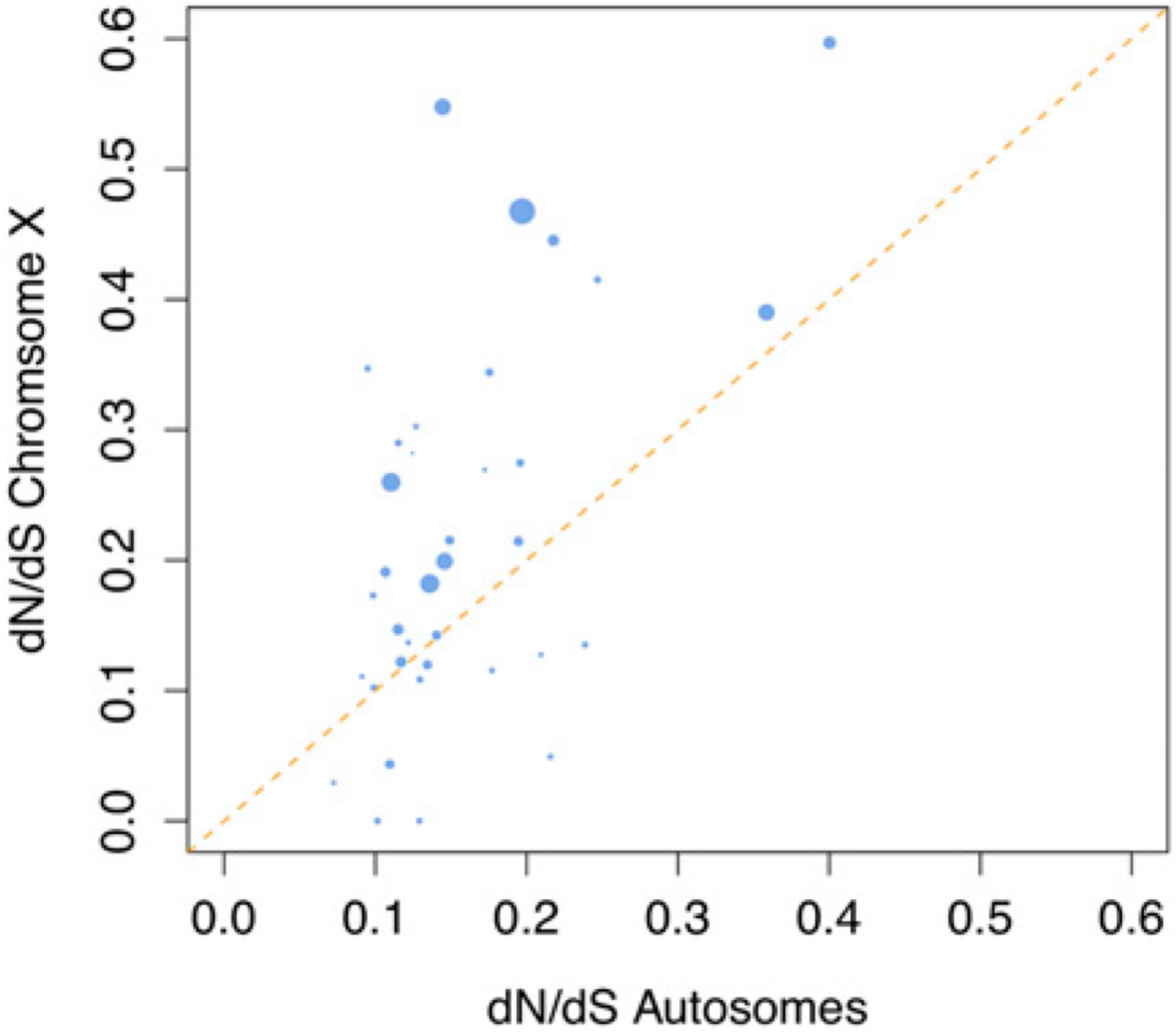


n=37 Gene Ontologies (GOs) and the speed of evolution (dN/dS) is measured on autosomes and X for all genes in a GO

Observation: **27 GOs** have a higher dN/dS on the X compared to Autosomes

Data: Hvilsom et al 2012 **Extensive X-linked adaptive evolution in central chimpanzees,** PNAS.

More about the X …A versatile use of the binomial test



n=37 Gene Ontologies (GOs) and the speed of evolution (dN/dS) is measured on autosomes and X for all genes in a GO

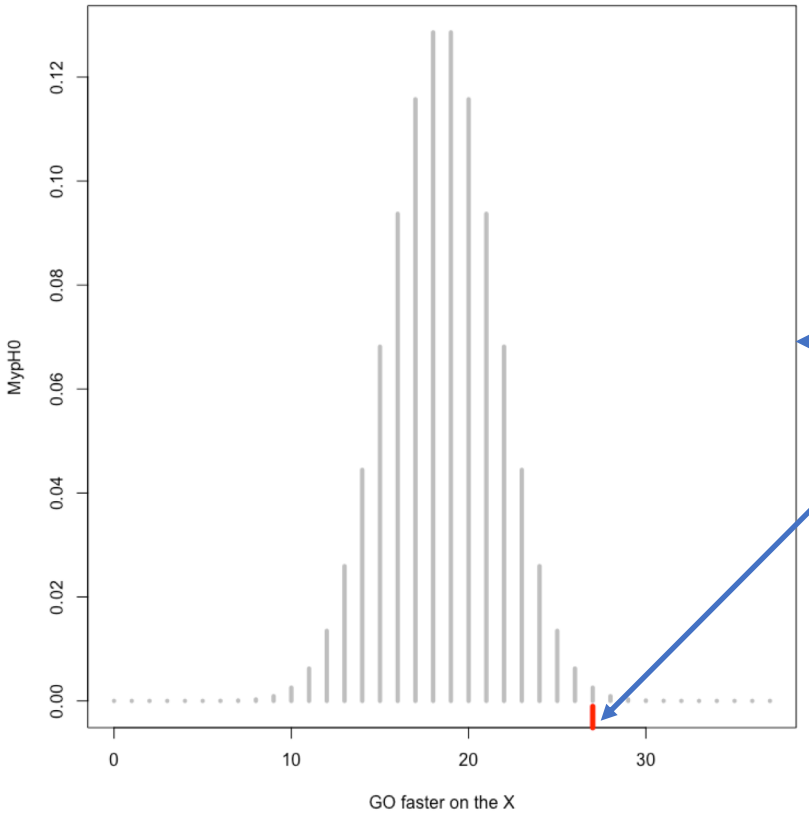
Observation: **27 GOs** have a higher dN/dS on the X compared to Autosomes

Is there statistical support for the fact that genes evolve faster on the X?

Our null: GO may predict how fast gene evolve BUT genes evolve at the same speed irrespective of being autosomal or X linked

Data: Hvilsom et al 2012 **Extensive X-linked adaptive evolution in central chimpanzees,** PNAS.

Doing the test from scratch in R …




```
xs=seq(from =0, to =37, by=1) ## X space

MypH0=dbinom(x = xs,size = 37,prob = 0.5) #H0

sum(MypH0) ## CHECK

> plot(xs,MypH0, col="grey", type="h", lwd=4,
xlab="GO faster on the X")

>rug(27, col="red", lwd=5) # the data


> 1- pbinom(q = 27,prob = 0.5,size = 37) #  H0 tail
1 sided

[1] 0.001281604
```

The binomial distribution can be used to test two types of null hypotheses:

For **proportions in a sample**

H0 $p=p_0$

Ha $p=p_a$

(there is an infinity of potential Ha)

Use counts as test statistic to decide whether p_0 makes the data plausible / implausible

Get a p-value and take a decision based on choice type I error

For **paired** observations

H0 is ”no differences between sexes / treatments / chromosomes”

Ha: there is a trend

- We ”control ” biological variation
- We just look qualitatively at differences

Use counts as test statistic to decide whether it is equally likely to be bigger/smaller or whether there is a trend

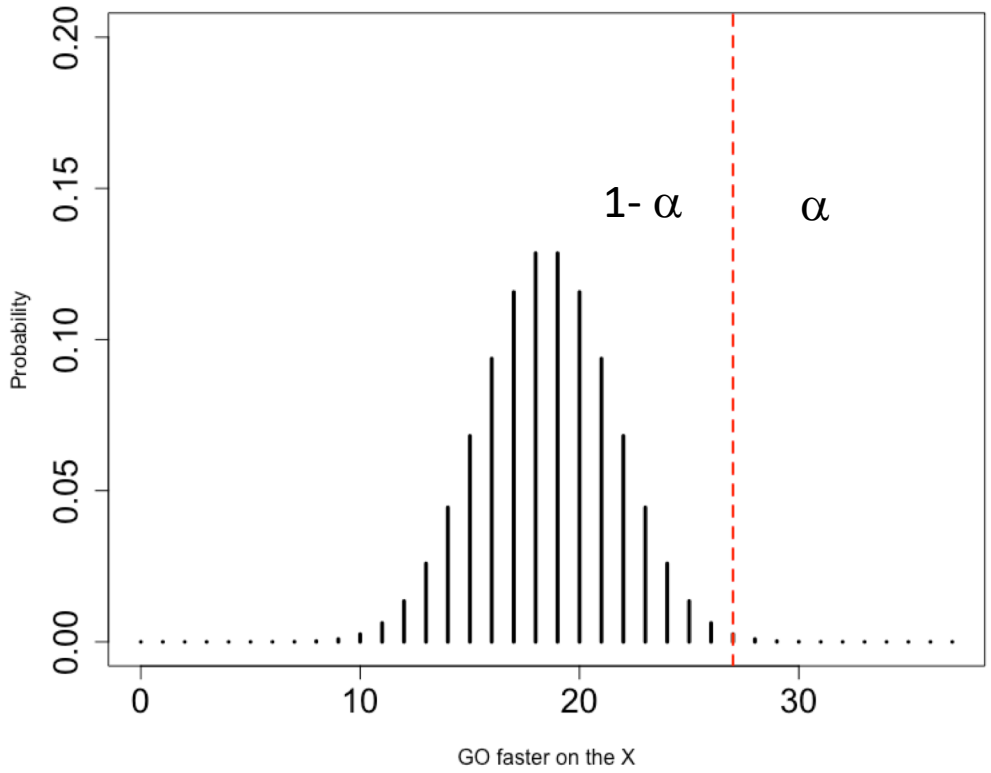
Get a p-value and take a decision based on choice type I error

Type I and Type II errors …

	H0 is TRUE	**H0 is False**
DECISION (based on $\mathcal{D}$)		
Reject H0	Type I error (α)	Correct
Do not reject H0	Correct	Type II error (β)

Choosing a significance level (α) for a P-value sets the type I error rate at α.
Varying α will have consequences for the rate of type II error and statistical power
Power = 1 – β
IF your test is well behaved/ calibrated and p-value well calculated THEN you will reject Wrongly H0 a proportion α of the time when making a test

Type I error



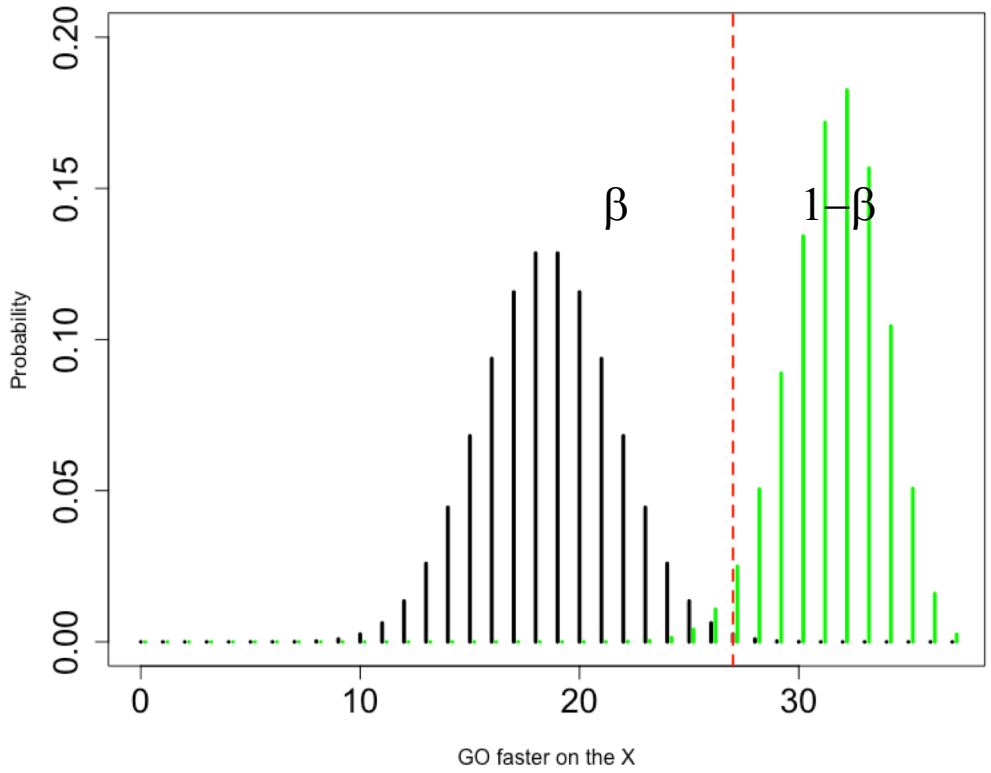
Type I error:

in principle you control that ☺

By definition, it ‘s the fraction α of datasets where you will reject H0, although the data is coming from the null

It is the threshold you choose for p values to ”declare significance”

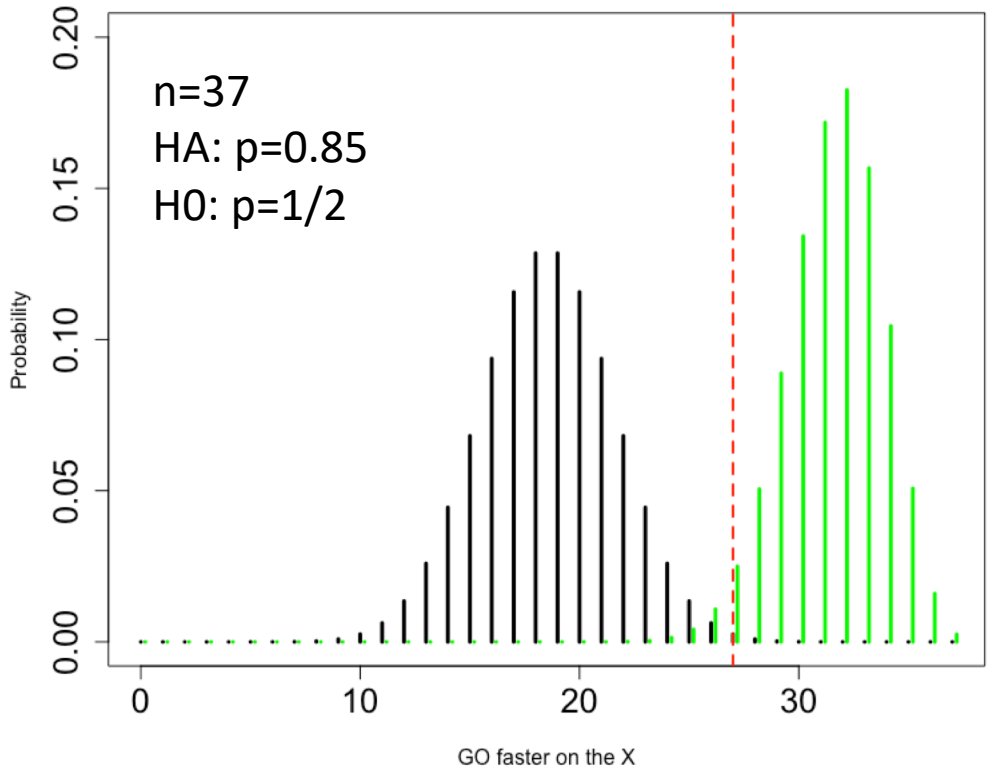
Type I and Type II error

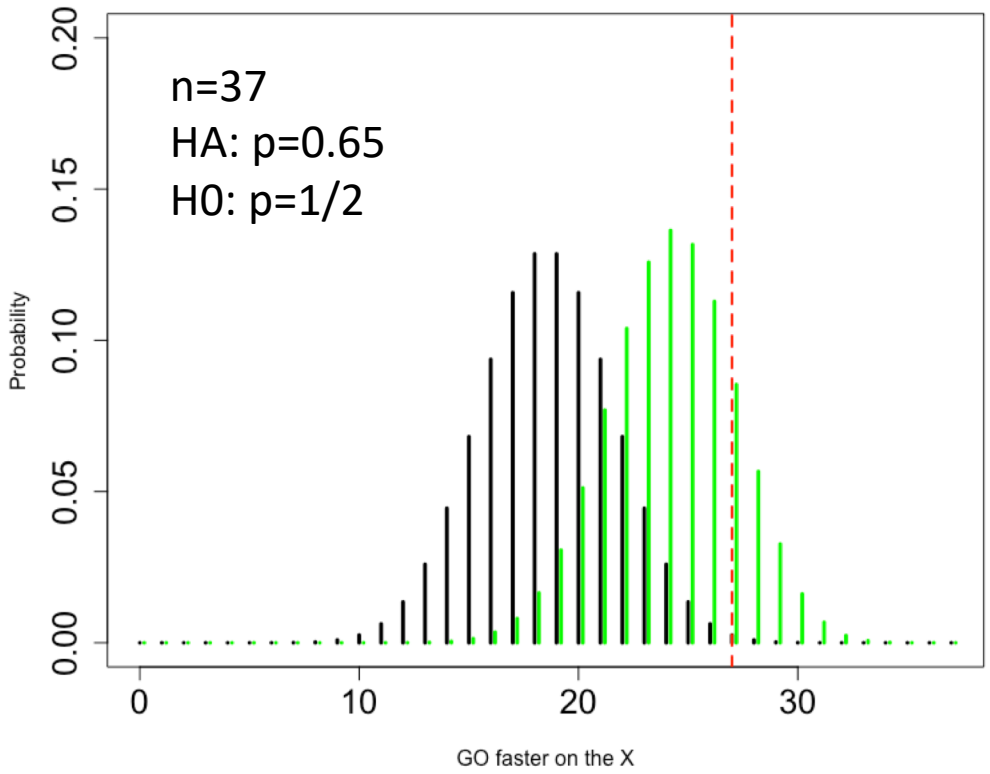


Statistical Power to reject H0, is the fraction of datasets obtained under the alternative Ha that will make you reject the null.

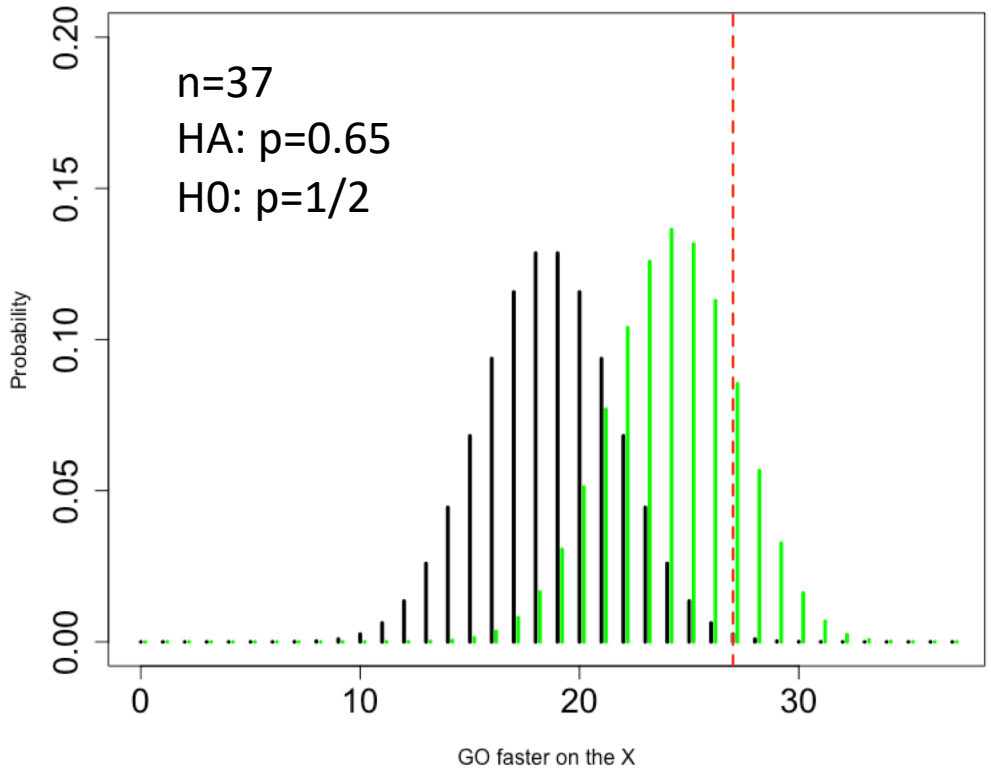
What is the power here ?

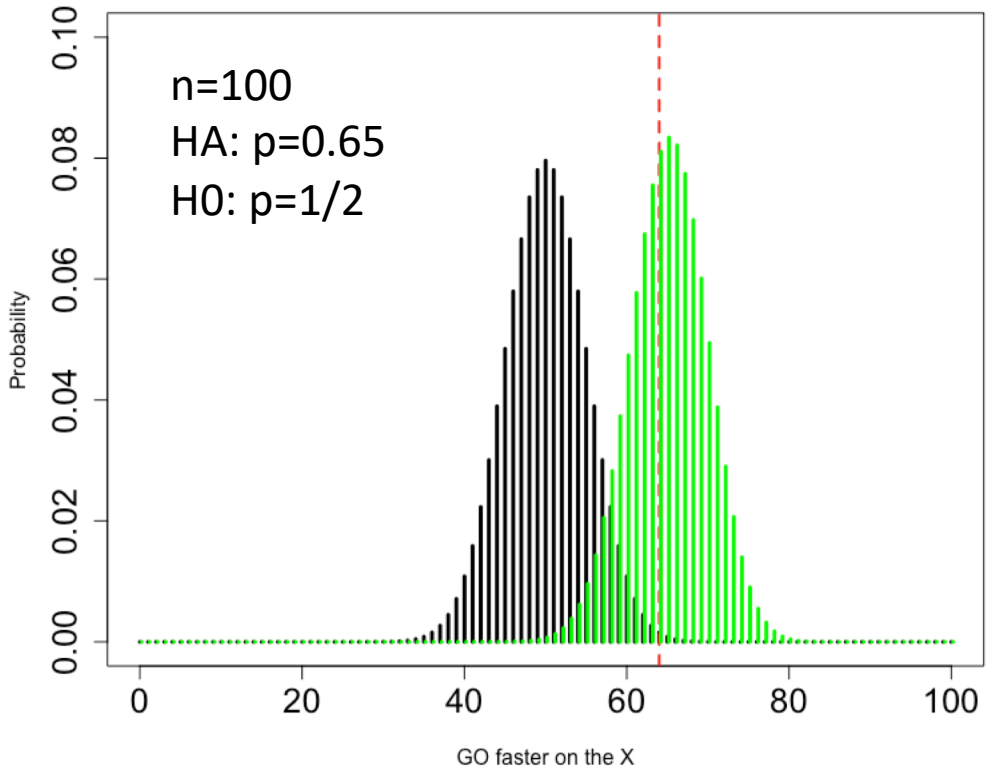
For identical sample size (n=37), power greatly depends on Ha you choose



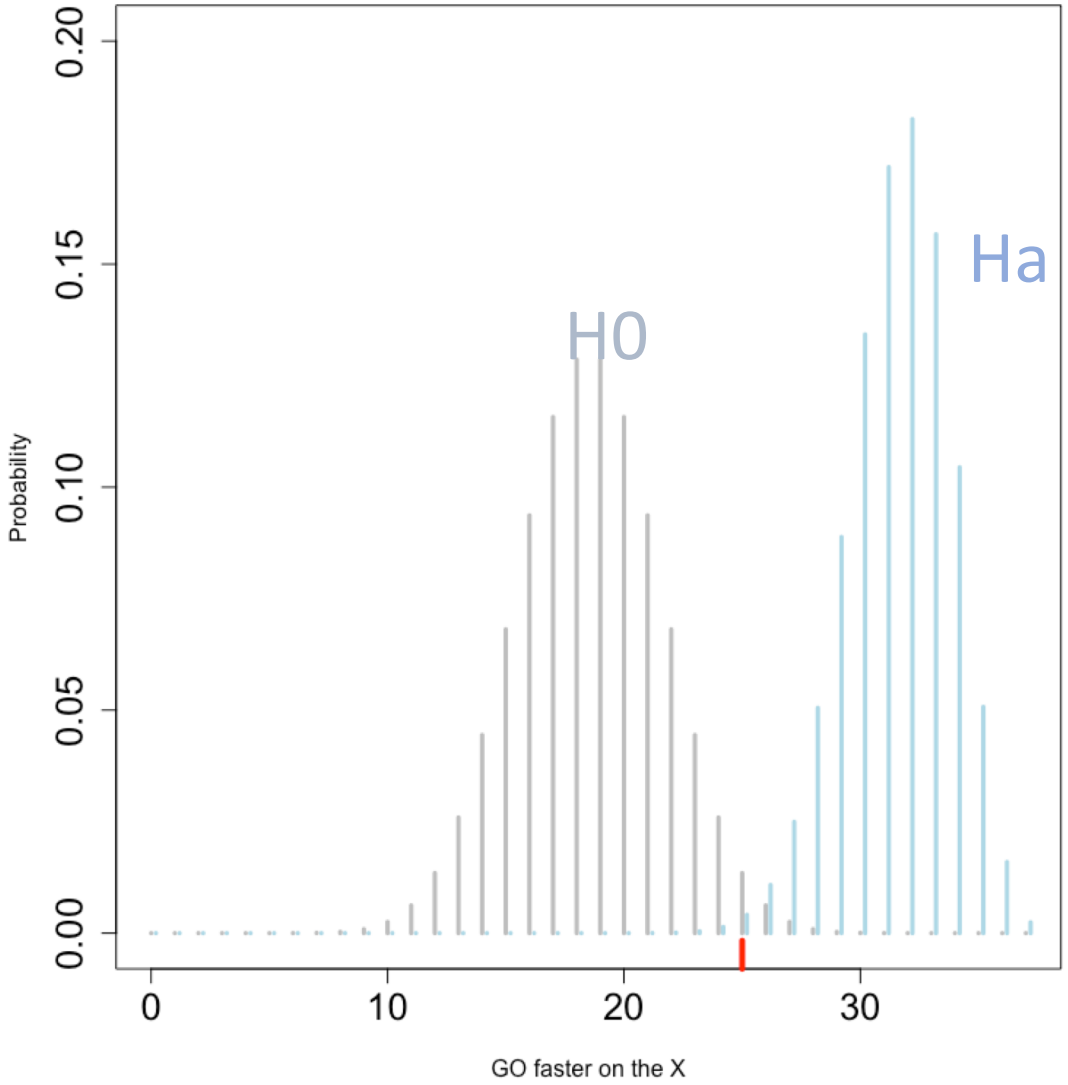


Increasing sample size helps (a lot): imagine we had 100 GO categories instead of 37





RECAP: Power, type I and type II errors:



Power studies ask : will I be able to reject H0 if the data is coming from Ha ?

We need to specify Ha

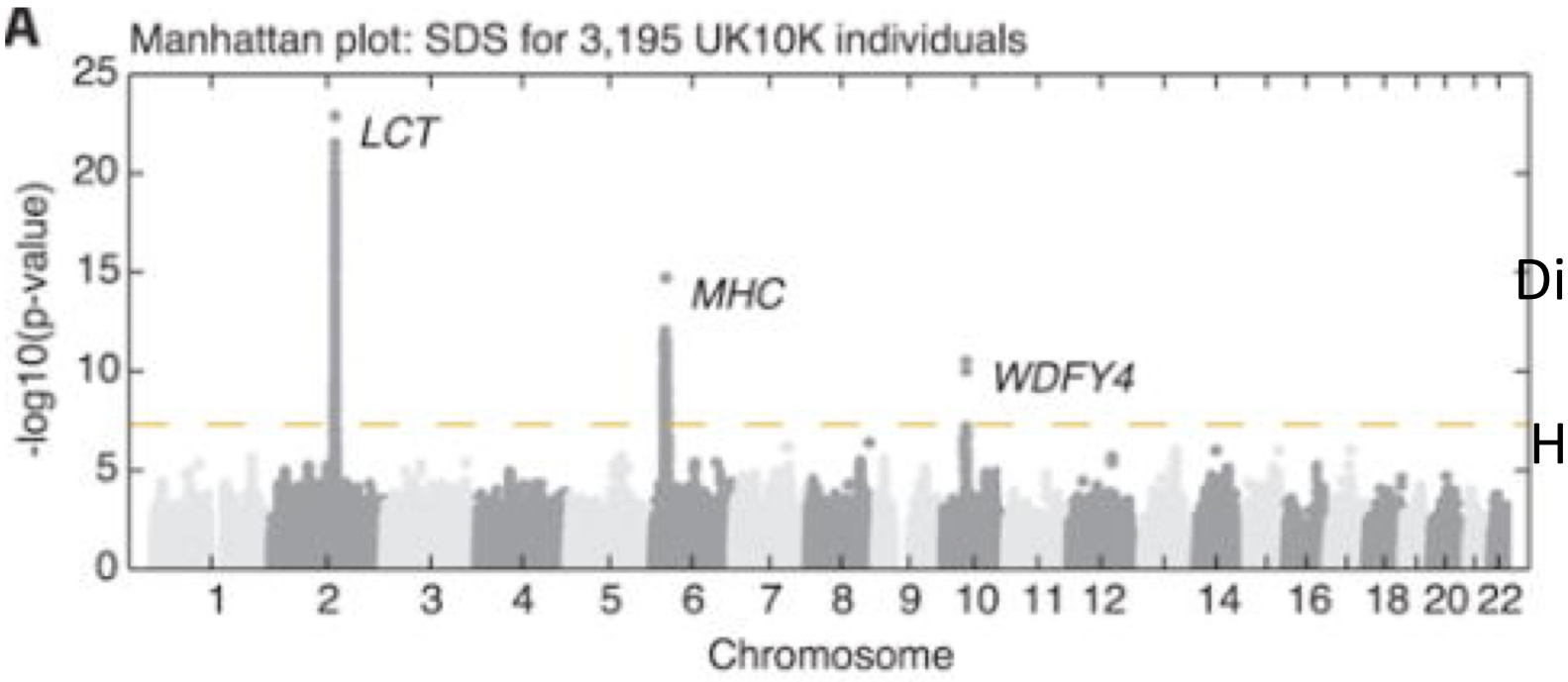
So we need a range of data values that may happen under a very specifc alternative

Here we use a different probability distribution as the way to define a precise alternative hypothesis

Genomic studies → many tests on a dataset

”We used the set of 3195 genomes to compute SDS for 4.5 million autosomal SNPs with minor allele frequency >5%.” Science. 2016 Nov 11; 354(6313): 760–764.
doi: 10.1126/science.aag0776

Big data → A dillemma



Many tests

A few ”discoveries”

Did we get these ”discoveries” just by chance?

How to make sure the one we flag as ”significant” are real?

Type I and type II errors dilemmas

Gene expression differences

10, 000 genes are tested for differences in gene expression between normal and cancer tissues.

Given the test and our sample size we ”know” (have calculated / simulated) that we have

Power 0.8 to reject H0, if we set α=0.05

Power 0.6 to reject H0, if we set α =0.01

We control type I error rates so …

What α should we choose to minimize ”false discovery rate” (FDR)

if we expect that

- 10 genes have genuine differences ?
- 100 genes have genuine differences?

FDR =

False positive/(False Positive + True Positive)

G Genes → G "decisions"

P_0: fraction of genes coming from H_0 (usually unkown)

$N_0 = Pi_0$ G genes are from H_0

$N_A = (1-Pi_0)$ G genes are from H_A

Scenario 1: 1-p0 = 0.001 G=10000

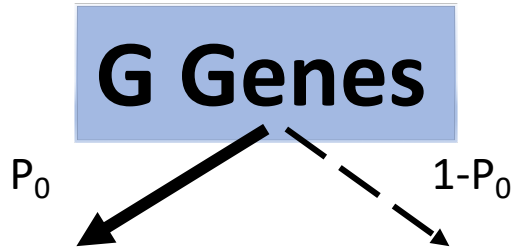
Scenario 2: 1- p0 =0.01 G=10000

Fact : FP and TP are random variables

TP and FP have a binomial distribution (indep tests, fixed prob of something happening)

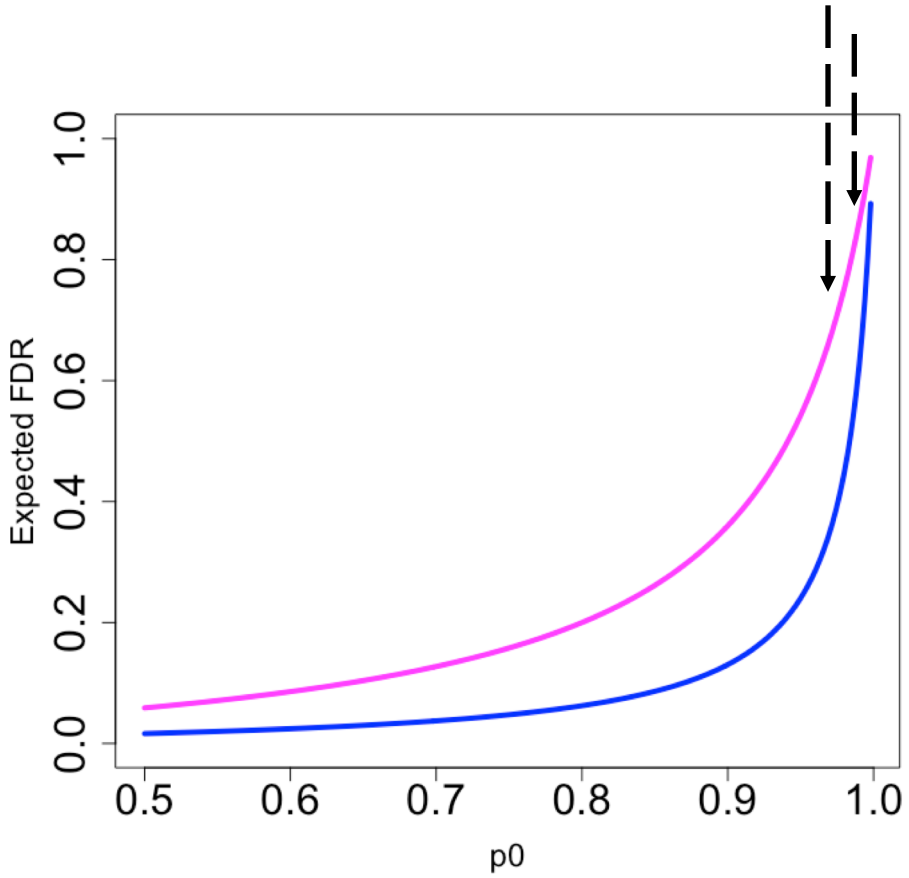
$$E(FP) = \alpha * p0 * G$$

$$E(TP) = (1-\beta)(1-p0) * G$$



	H0 is TRUE N_0 genes	H0 is False N_A genes
DECISION (based on $\mathcal{D}ata$)		
Reject H0	**FALSE POSITIVE** (FP) Type I error (α)	**TRUE POSITIVE** (TP) ($1-\beta$)
Do not reject H0	TRUE NEGATIVE ($1-\alpha$)	FALSE NEGATIVE Type II error (β)

A graphical overview



Here in both cases ”blue” wins

The ”blue” strategy dominates the ”pink” one

→Being conservative ”pays” ☺

→We are drowned in many genes so we want to drastically limit type I error

→Note: every situation will change the way α and $1-\beta$ co-vary

Power 0.8 to reject H0, if we set α=0.05

Power 0.6 to reject H0, if we set α =0.01

What remedies for multiple testing

- Do less tests (if possible)
- Correct p-values
- Increase sample size (if feasible)

Bonferonni correction

Control false discovery rates

Goals for this week

Stats concepts

- Binomial random variable
- E() and V() of the mean proportion
- type I error and power
- Binomial tests for proportions

R skills

- summarizing data with means (medians) and proportions
- ggplot to vizualize scatter plots and trends (outlook to regression)
- binomial tests in R
  - from scractch
  - binom.test()