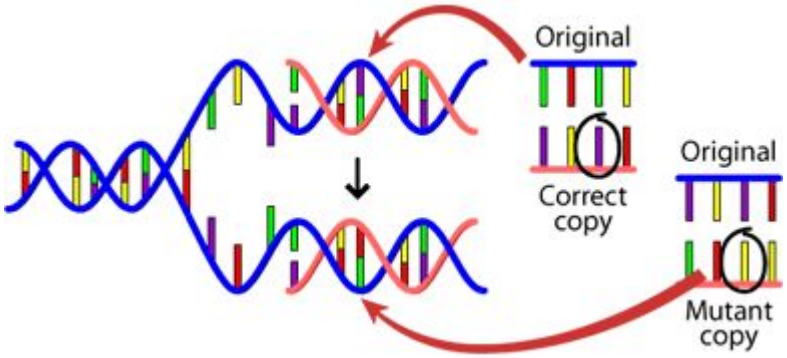
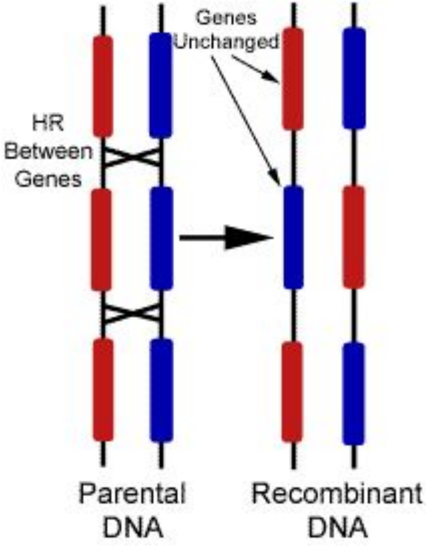
Week 05: Quantitative genetics

- Genome-wide association studies
    - Complex traits
    - Statistical inference, P-values, & Multiple hypothesis testing
    - Regularized linear regression
    - Polygenic risk score

Genetic variation





Single Nucleotide Polymorphisms (SNPs)

Insertions

Deletions

Copy Number Variants (CNVs)

- Duplications & deletions

@genomicsedu

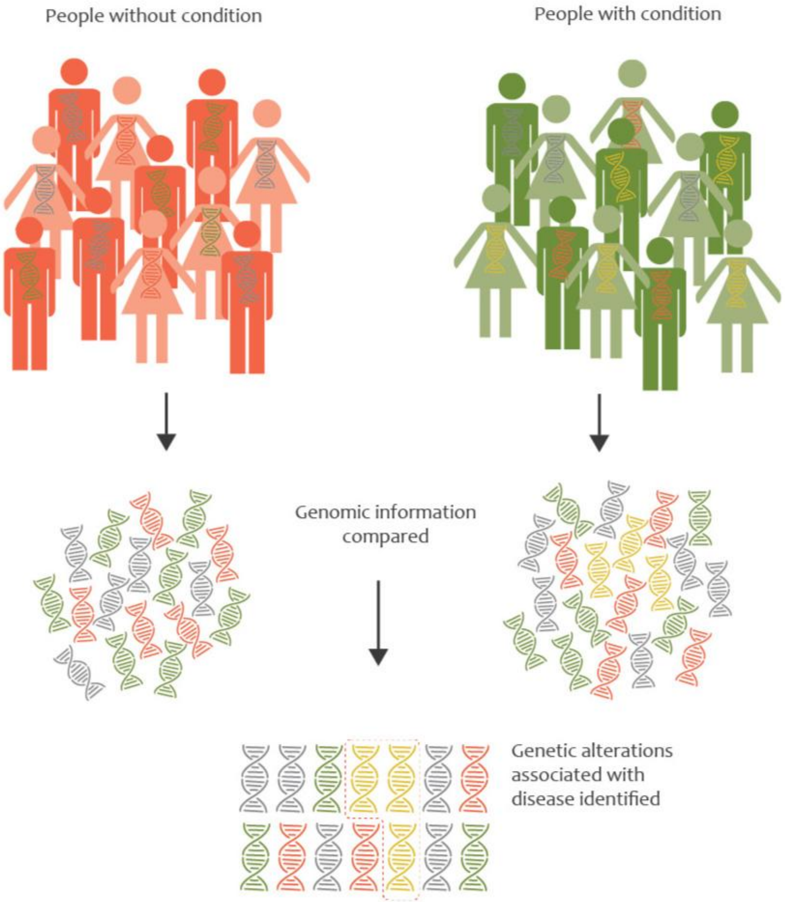
Complex traits and diseases

What factors contribute to a particular trait or the risk of getting a particular disease?

- Genetic factors (numerous)
- Other biological factors: age, sex, ethnicity
- Environmental factors (e.g. geography, nutrition)
- Interaction between genome and environment
  - Phenotypic Variation = G + E + GxE

How do you quantify how much the genome actually contributes?

Genome-wide Association Study (GWAS)


People without condition
People with condition
Genomic information compared
Genetic alterations associated with disease identified

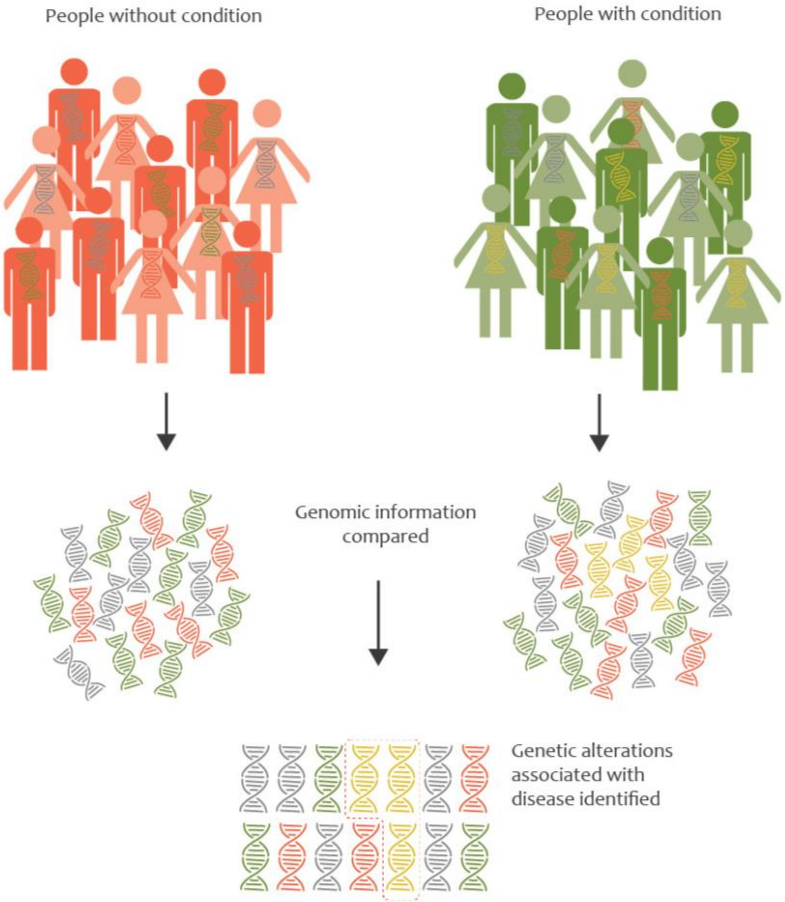
Still expensive to sequence entire genome.

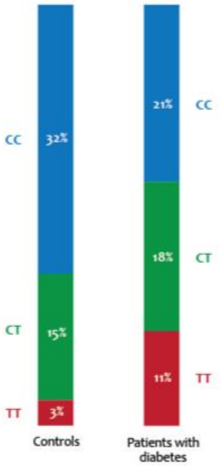
Focus on only a small part of the genome (SNPs) that are common and might contribute to variation.

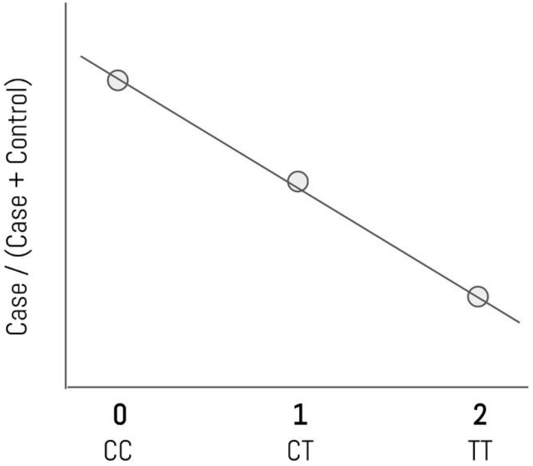
- About 5–10 million SNPs in the human genome.
- Use a SNP array – a small chip that has DNA probes that is complementary to regions in the genome that have SNPs.

Visscher (2017) AJHG; @genomicsedu

Genome-wide Association Study (GWAS)


People without condition
People with condition
Genomic information compared
Genetic alterations associated with disease identified





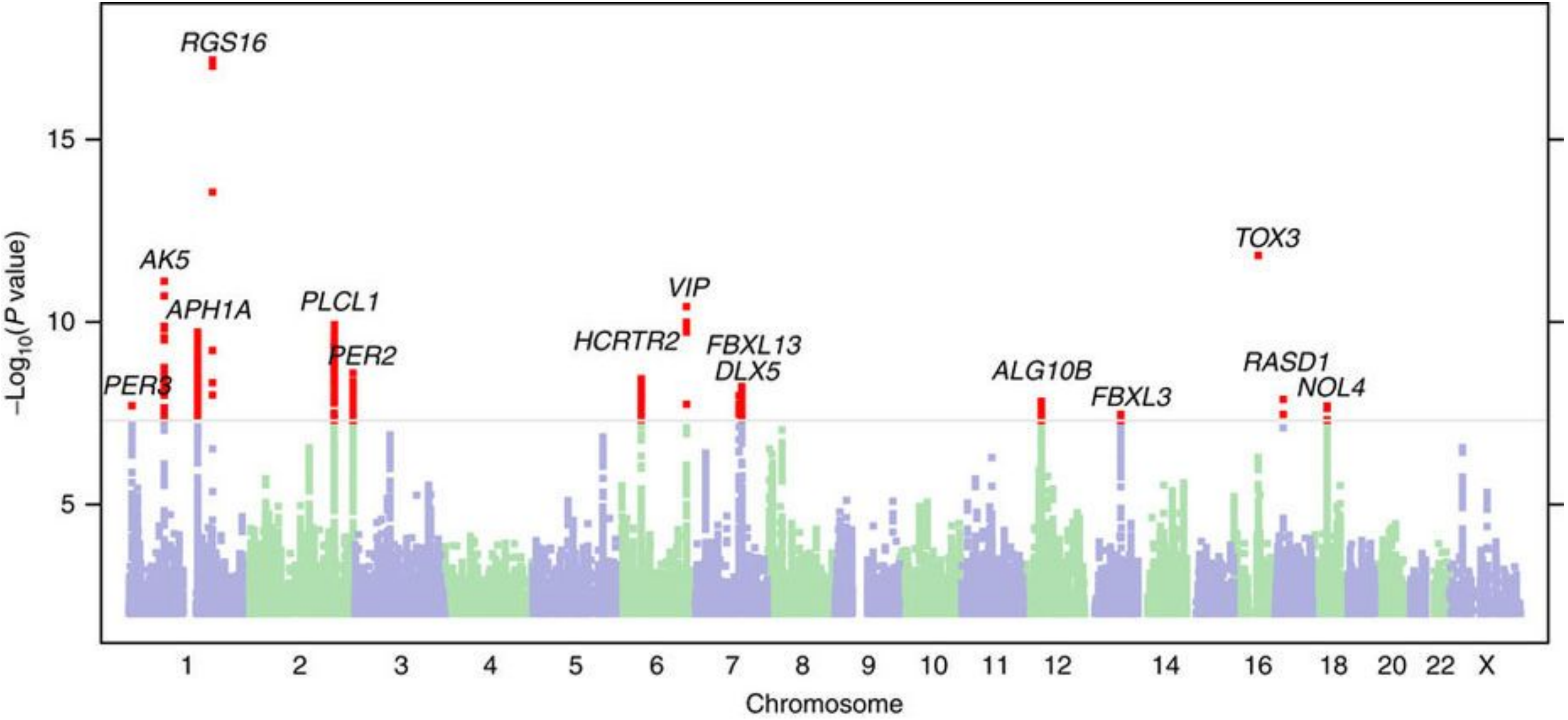
A C/T SNP from a hypothetical GWAS for type 2 diabetes

- Increase in freq of T allele in patients w/ diabetes compared to controls.
- We know where this SNP is on the genome → study surrounding sequence

Visscher (2017) AJHG; @genomicsedu

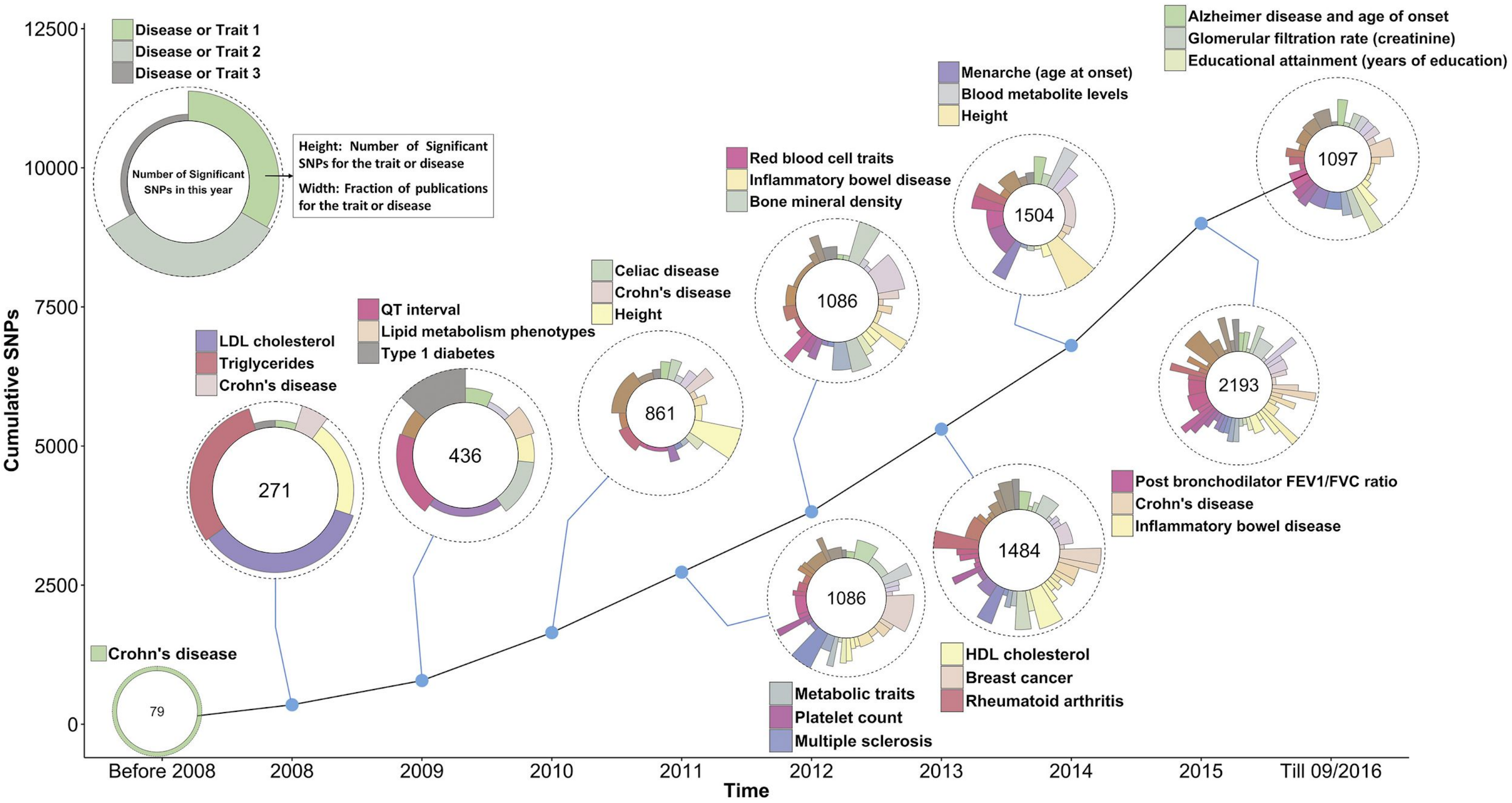
Results of a GWAS

GWAS of 89,283 individuals identifies genetic variants associated with… being a morning person!



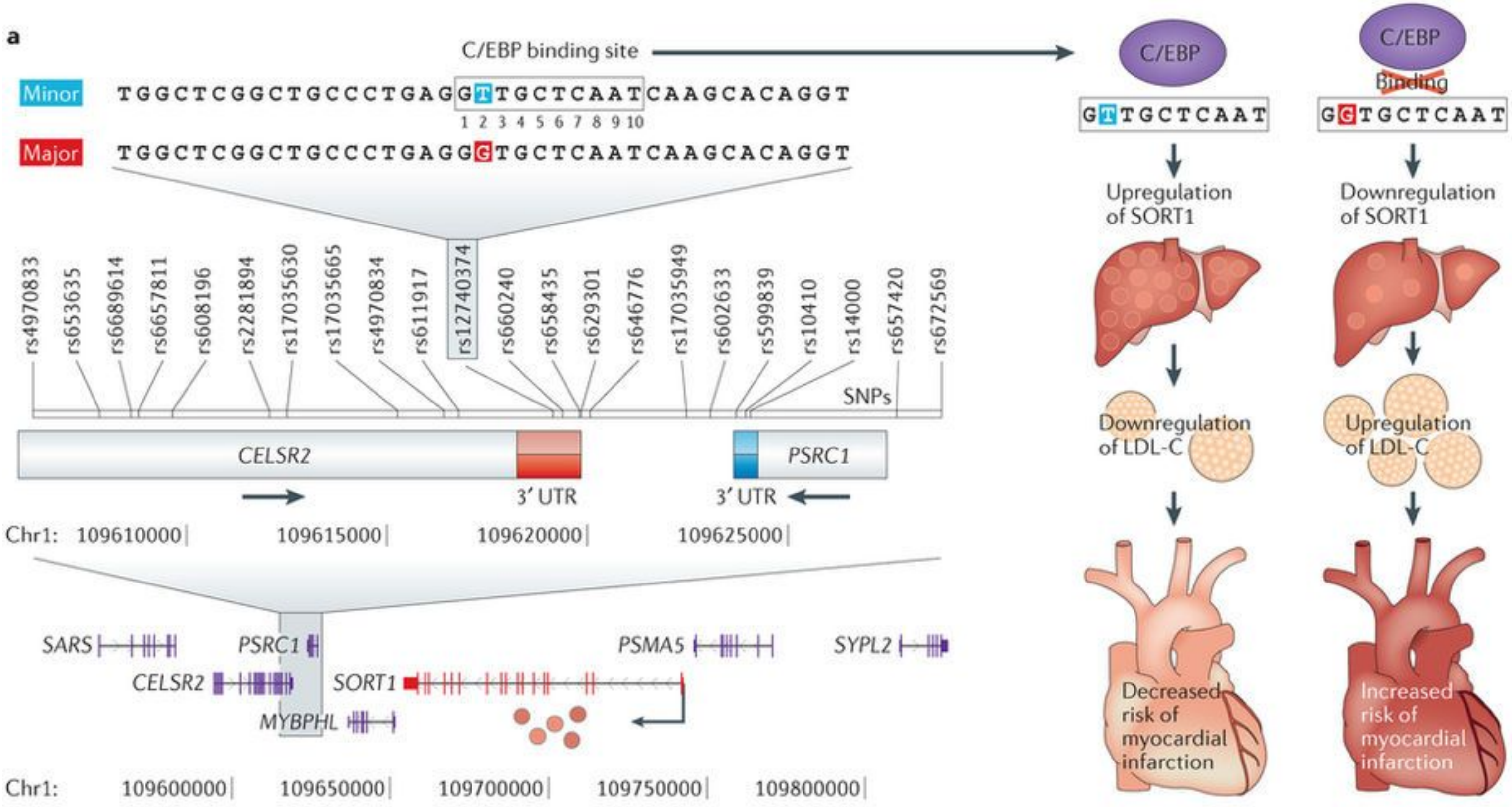
Hu (2016) Nat. Comm.

GWAS – Timeline of discoveries



Visscher (2017) AJHG

GWAS – Examples



Albert & Kruglyak (2015) Nat. Rev. Genet.

GWAS – Examples

Variation in the **nicotinic receptor** leads to higher levels of **lung cancer** in the developed world.

- This is *not* because the nicotinic receptor is directly involved in the molecular aspects of lung cancer.
  - Rather these variants make people get a bigger hit from nicotine.
  - I.e. *if* they start smoking, they are less likely to *stop* smoking (more smoke exposure).
- So, this variant is causally involved in lung cancer.
  - I.e. if one has it, their odds are fundamentally higher.
- However, the mechanism will not be clear if we didn't know about nicotine from other studies.
- Smoking exposure is the **main cause** & this variant in the nicotine receptor is a **modifier**.

Statistical hypothesis testing

1. **Decide on the effect** that you are interested in, design a suitable experiment or study, pick a data summary function and test statistic.

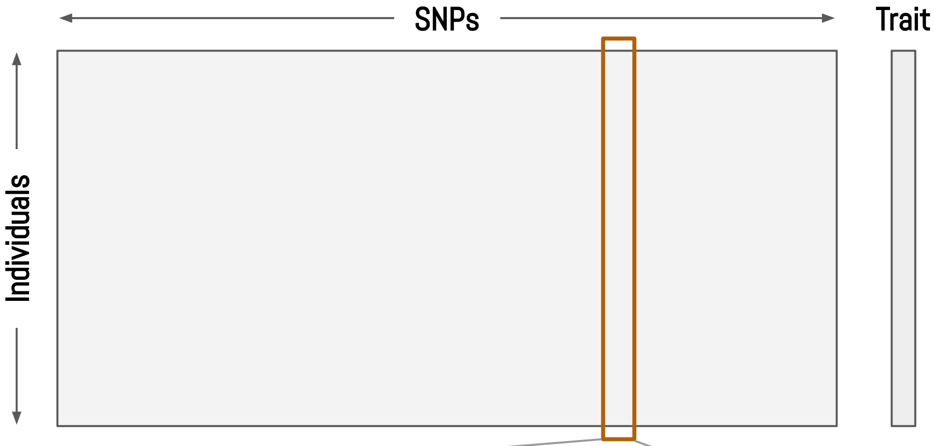
2. **Set up a null hypothesis**, which is a simple, computationally tractable model of reality that lets you compute the null hypothesis.

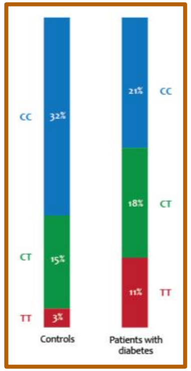
3. **Decide on the rejection region**, i.e., a subset of possible outcomes whose total probability is small.

4. **Do the experiment** and collect the data, compute the test statistic.

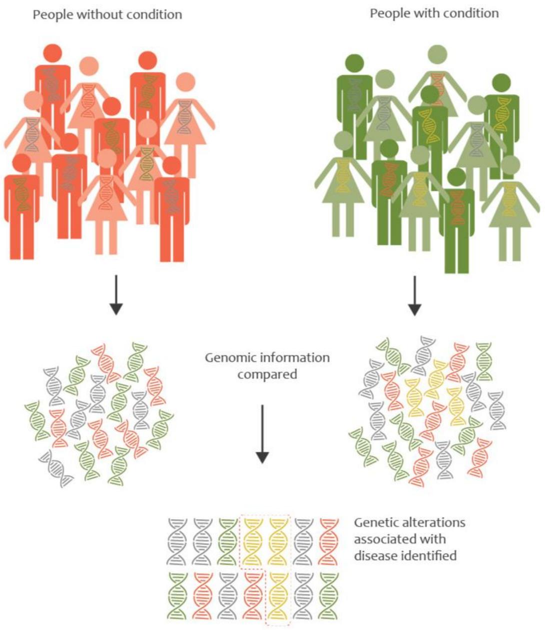
5. **Make a decision**: reject the null hypothesis – i.e. conclude that it is unlikely to be true – if the test statistic is in the rejection region.

Statistical hypothesis testing for GWAS





	$X_n = 0$	$X_n = 1$	$X_n = 2$	**Totals**
$Y = 0$	O_{00}	O_{01}	O_{02}	$O_{0.}$
$Y = 1$	O_{10}	O_{11}	O_{12}	$O_{1.}$
Totals	$O_{.0}$	$O_{.1}$	$O_{.2}$	S


People without condition
People with condition
Genomic information compared
Genetic alterations associated with disease identified

Pearson's χ^2 test

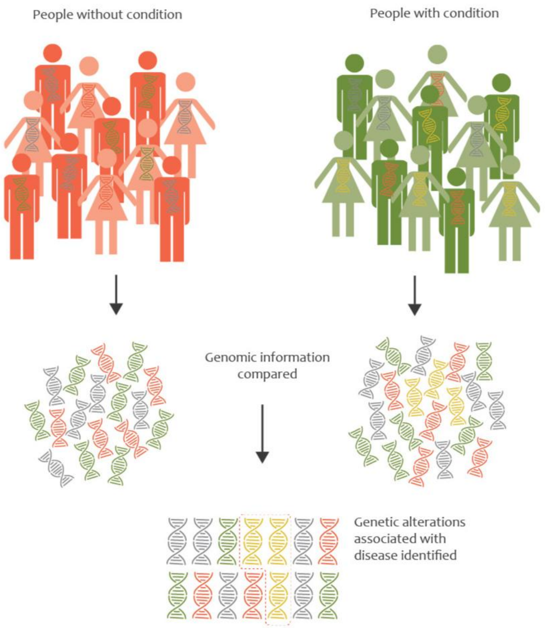
$$X^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

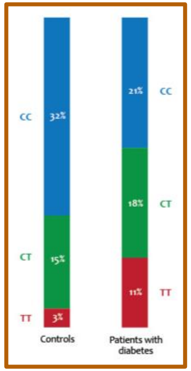
Statistical hypothesis testing for GWAS

Consider two competing hypotheses for a given SNP:

- **Null hypothesis**: the frequency of the SNP in the cases is the same as that in controls.
- **Alternative hypothesis**: the frequencies are different.

There's always some difference → Is it significant difference?


People without condition
People with condition
Genomic information compared
Genetic alterations associated with disease identified



	$X_n = 0$	$X_n = 1$	$X_n = 2$	**Totals**
$Y = 0$	O_{00}	O_{01}	O_{02}	$O_{0.}$
$Y = 1$	O_{10}	O_{11}	O_{12}	$O_{1.}$
Totals	$O_{.0}$	$O_{.1}$	$O_{.2}$	S

Pearson's χ^2 test

$$X^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Statistical hypothesis testing for GWAS

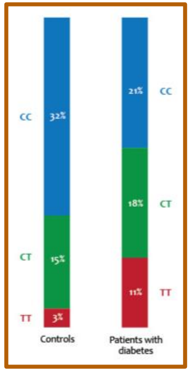
Consider two competing hypotheses for a given SNP:

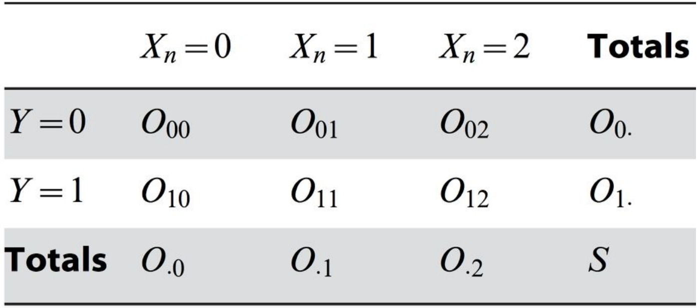
- **Null hypothesis**: the frequency of the SNP in the cases is the same as that in controls.
- **Alternative hypothesis**: the frequencies are different.

How is this question typically answered?

Calculate the p-value?

There's always some difference ⟶ Is it significant difference?





	$X_n = 0$	$X_n = 1$	$X_n = 2$	**Totals**
$Y = 0$	O_{00}	O_{01}	O_{02}	$O_{0.}$
$Y = 1$	O_{10}	O_{11}	O_{12}	$O_{1.}$
Totals	$O_{.0}$	$O_{.1}$	$O_{.2}$	S

Pearson's χ^2 test

$$X^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

What is the P-value?

The p-value is:

A. The amount of evidence that the SNP is associated with the trait/disease

B. The probability that the SNP is not associated

C. The probability that a SNP picked as associated is actually not

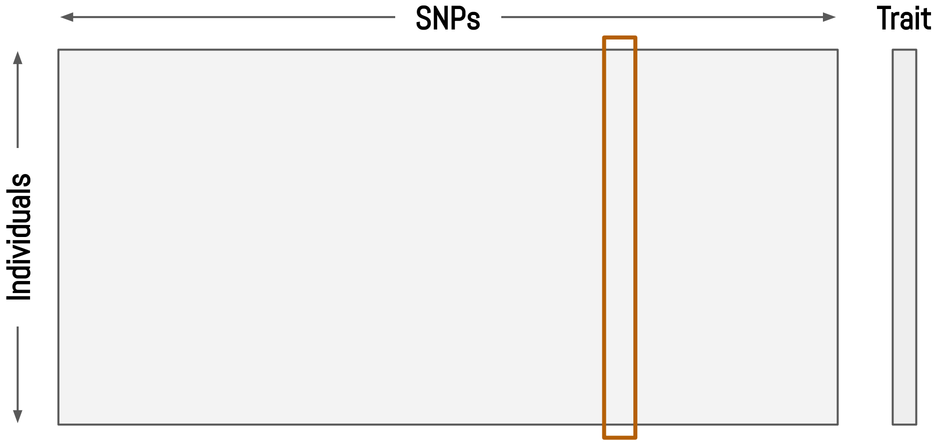
D. The strength of the SNP's effect on the trait/disease

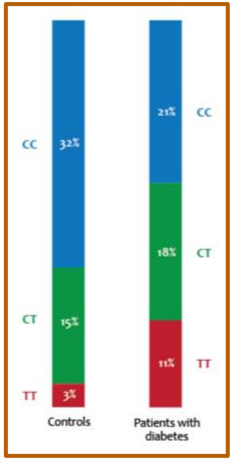
E. The probability that the outcome of the GWAS is important

The p-value is the probability that the study would have produced the observed outcome (or something more extreme) even if the SNP is not associated with the trait/disease.

How to calculate the P-value?

The p-value is the probability that the study would have produced the observed outcome (or something more extreme) even if the SNP is not associated with the trait/disease.





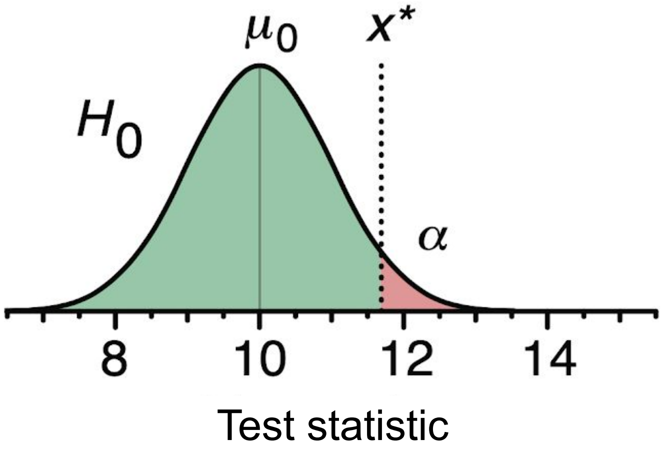
	$X_n = 0$	$X_n = 1$	$X_n = 2$	**Totals**
$Y = 0$	O_{00}	O_{01}	O_{02}	$O_{0.}$
$Y = 1$	O_{10}	O_{11}	O_{12}	$O_{1.}$
Totals	$O_{.0}$	$O_{.1}$	$O_{.2}$	S

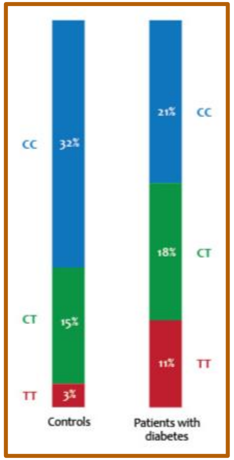
$$X^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

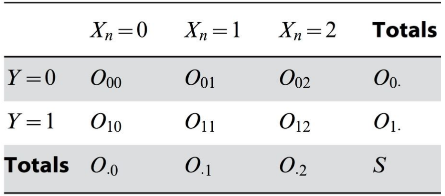
How to calculate the P-value?

The p-value is the probability that the study would have produced the observed outcome (or something more extreme) even if the SNP is not associated with the trait/disease.

The p-value is the area under the null distribution corresponding to outcome equal to or more extreme than the observed statistic.







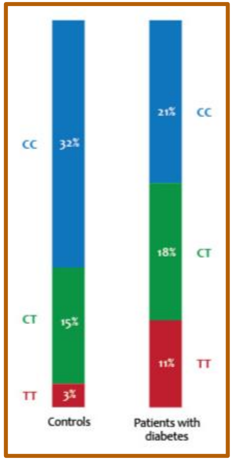
	$X_n = 0$	$X_n = 1$	$X_n = 2$	**Totals**
$Y = 0$	O_{00}	O_{01}	O_{02}	$O_{0.}$
$Y = 1$	O_{10}	O_{11}	O_{12}	$O_{1.}$
Totals	$O_{.0}$	$O_{.1}$	$O_{.2}$	S

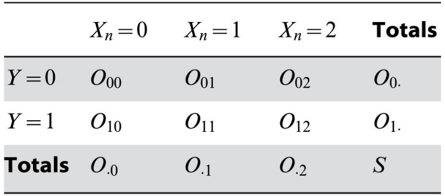
$$X^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

How to calculate the P-value?

The p-value is the probability that the study would have produced the observed outcome (or something more extreme) even if the SNP is not associated with the trait/disease.

1. Calculate the real test statistic.
2. Repeat the following 100,000 times to set up the null hypothesis for this test statistic:
   - ○ Randomly assign individuals to groups.
   - ○ Record the test statistic of the permuted assignments.
3. Calculate the p-value of the real test statistic. [**How?**]

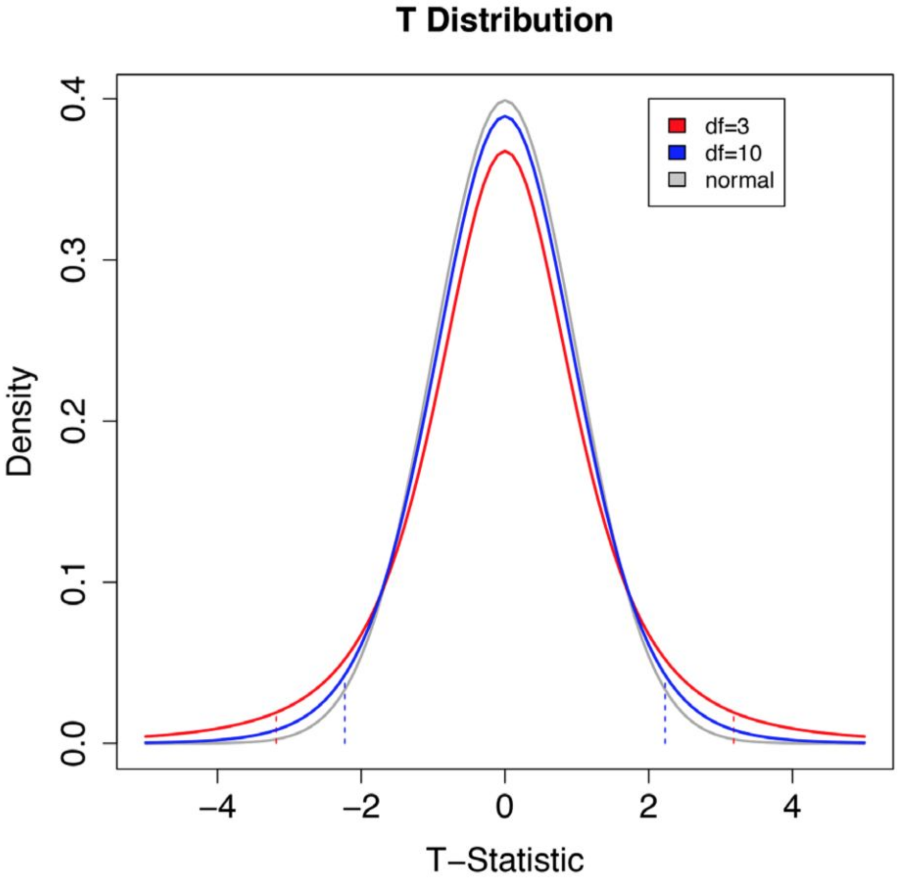




	$X_n = 0$	$X_n = 1$	$X_n = 2$	**Totals**
$Y = 0$	O_{00}	O_{01}	O_{02}	$O_{0.}$
$Y = 1$	O_{10}	O_{11}	O_{12}	$O_{1.}$
Totals	$O_{.0}$	$O_{.1}$	$O_{.2}$	S

$$X^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

How to calculate the P-value?



The p-value is the area under the null distribution corresponding to outcome equal to or more extreme than the observed statistic.

Student's one-sample test

$$t = \frac{\bar{x} - \mu_0}{\mathrm{SEM}}$$

Welch's two-sample test

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

P-value - History

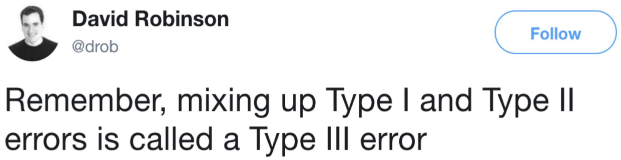
- Fisher (1920s):
    - Informal method to help interpret the data along with prior experience, domain knowledge, size of the effect, etc.

- Neyman & Pearson:
    - Control false positive rate at α, set by the experimenter based on what can be tolerated.
    - Formulate null and alternative hypothesis.
    - Reject null when $p < \alpha$.
        - The threshold $\alpha = 0.05$ is merely a convention.

Type I & type II errors

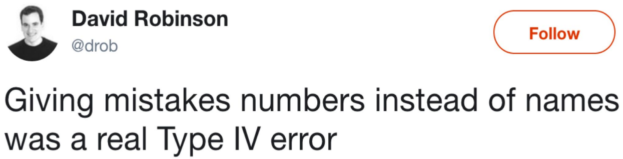
Choosing $p < \alpha$ controls Type I error at α.

- Type I error: False-positive rate (α)
- Type II error: False-negative rate (β)
- Remember the story of the boy that cried wolf!



David Robinson
@drob

Remember, mixing up Type I and Type II errors is called a Type III error



David Robinson
@drob

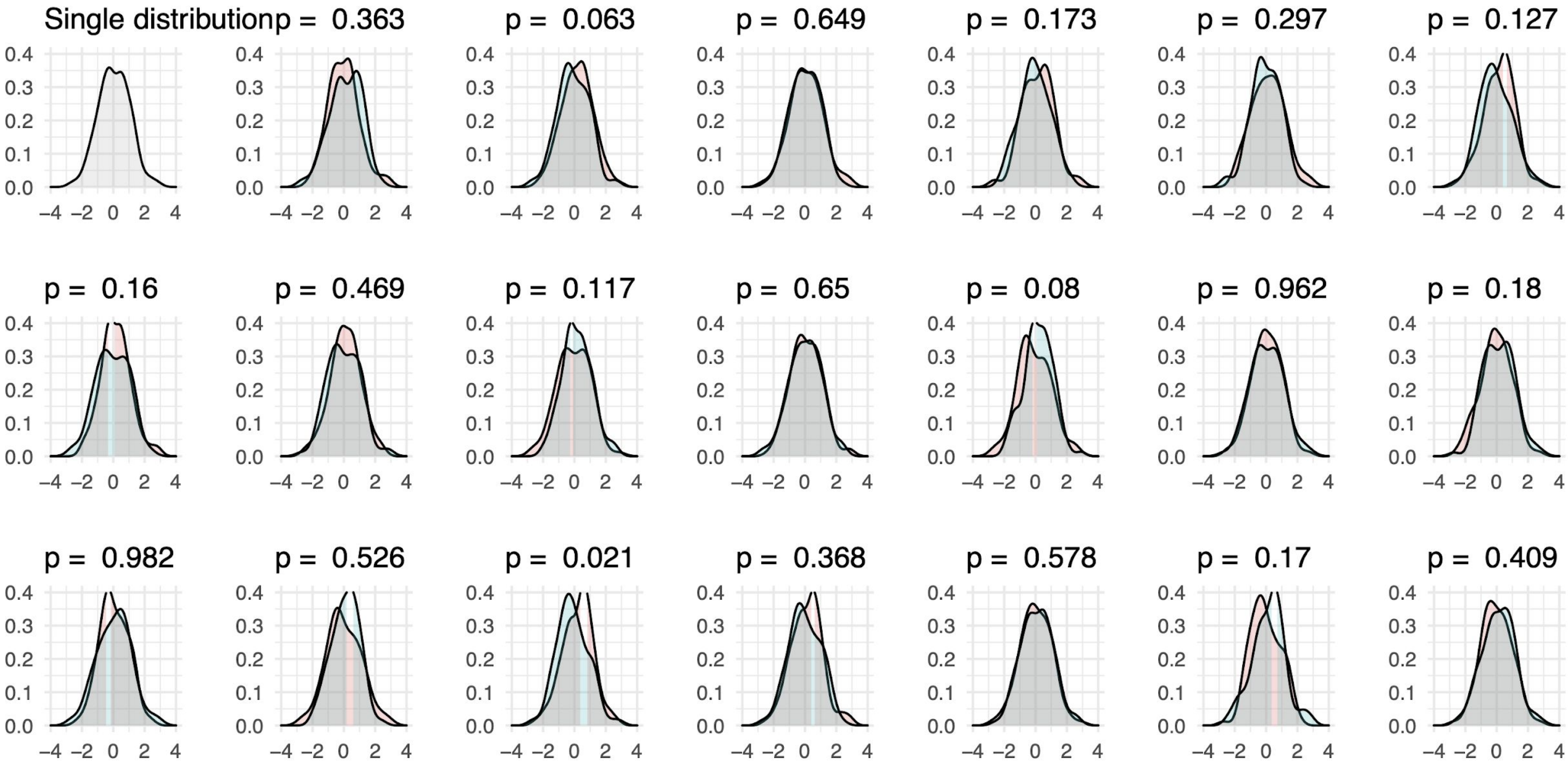
Giving mistakes numbers instead of names was a real Type IV error

P-value depends on multiple factors

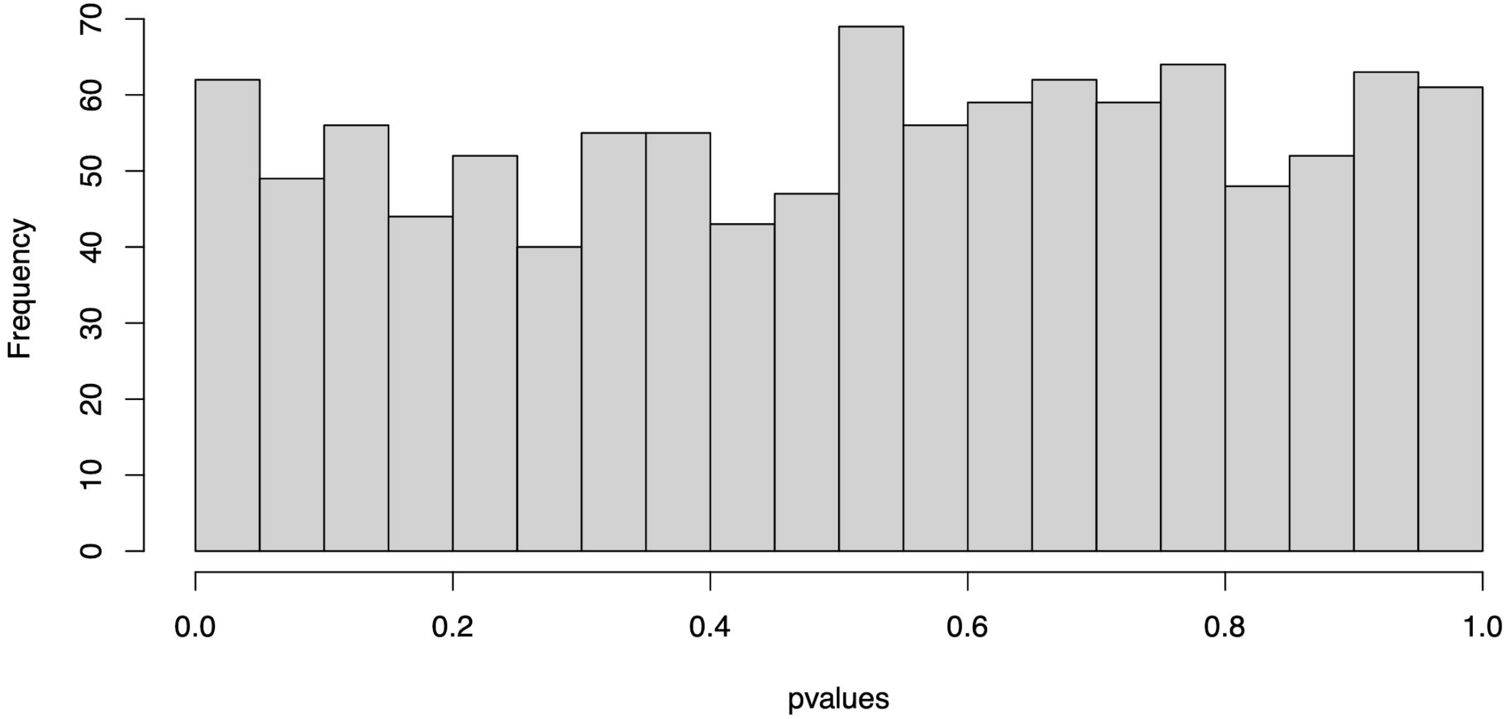
P-values are dependent on:

- Size of the effect (effect size)
- Variance within each group
- Sample size
- The underlying experimental design & the null hypothesis (need not always be random chance).
    a. Conversely, two completely different experiments can give same data but end up very different p-values.
        - 3 out of 9: Binomial p-value = 0.073
        - 3 out of 9: Neg. Binomial p-value = 0.033.

Distribution of p-values under the null hypothesis

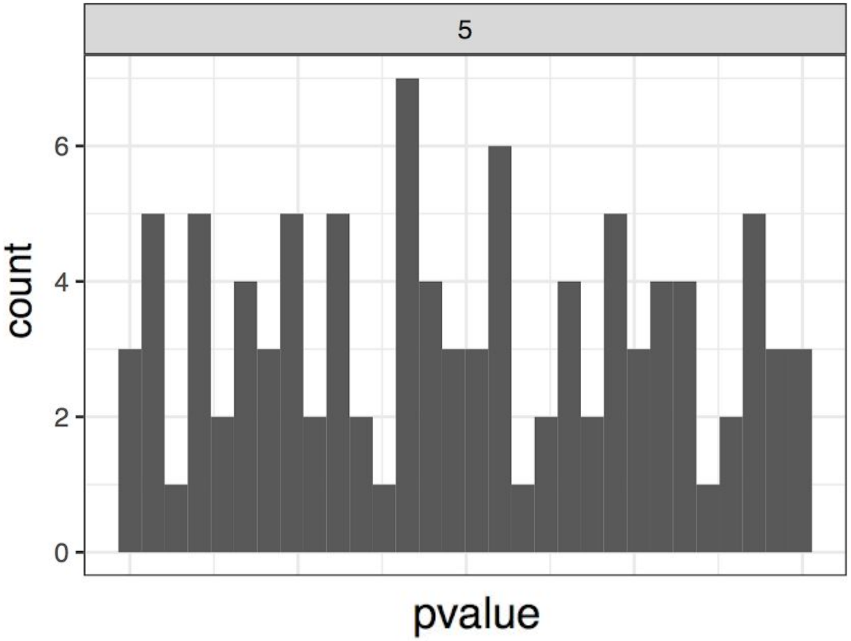


Distribution of p-values under the null hypothesis



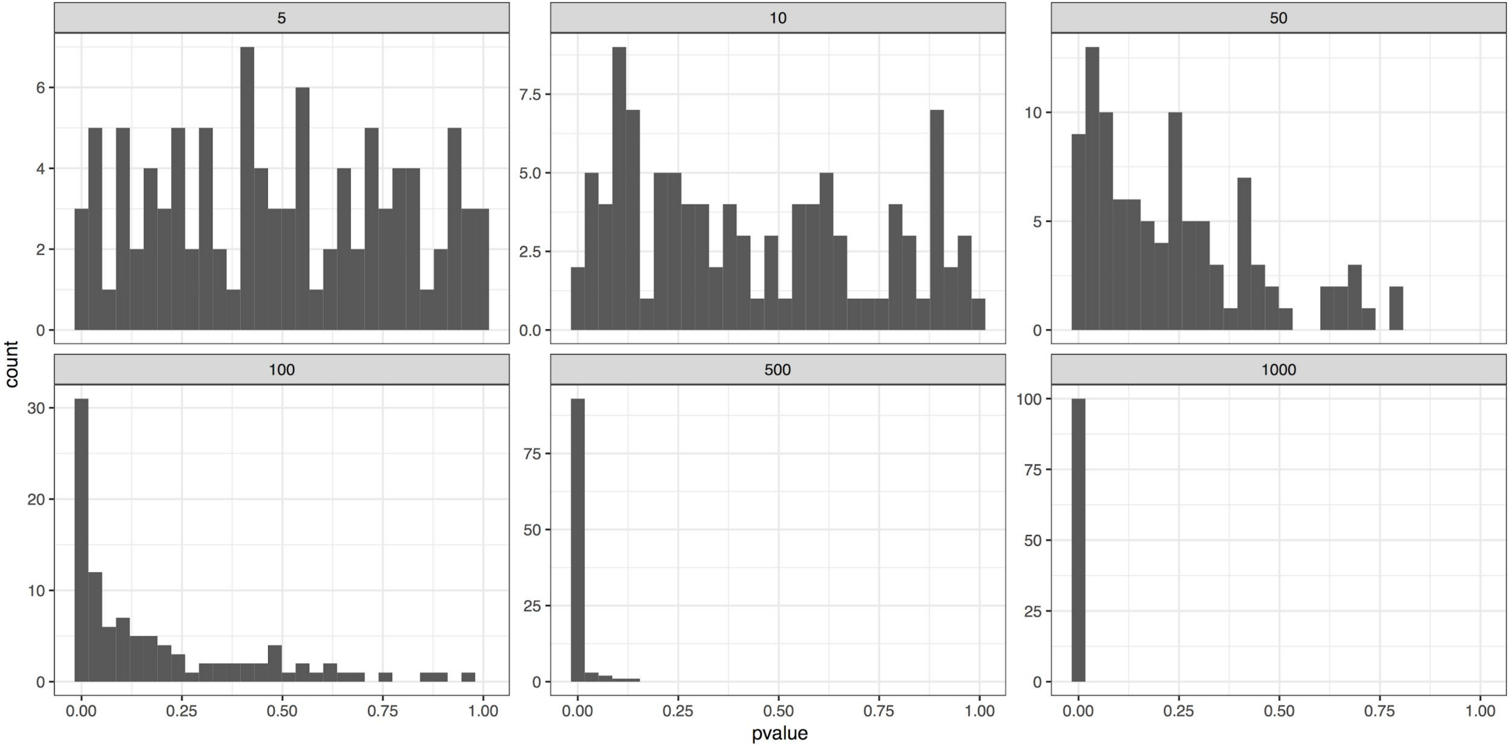
P-value depends on multiple factors

- P-values are dependent on: sample_size (effect_size = 0.25, std_deviation = 1)



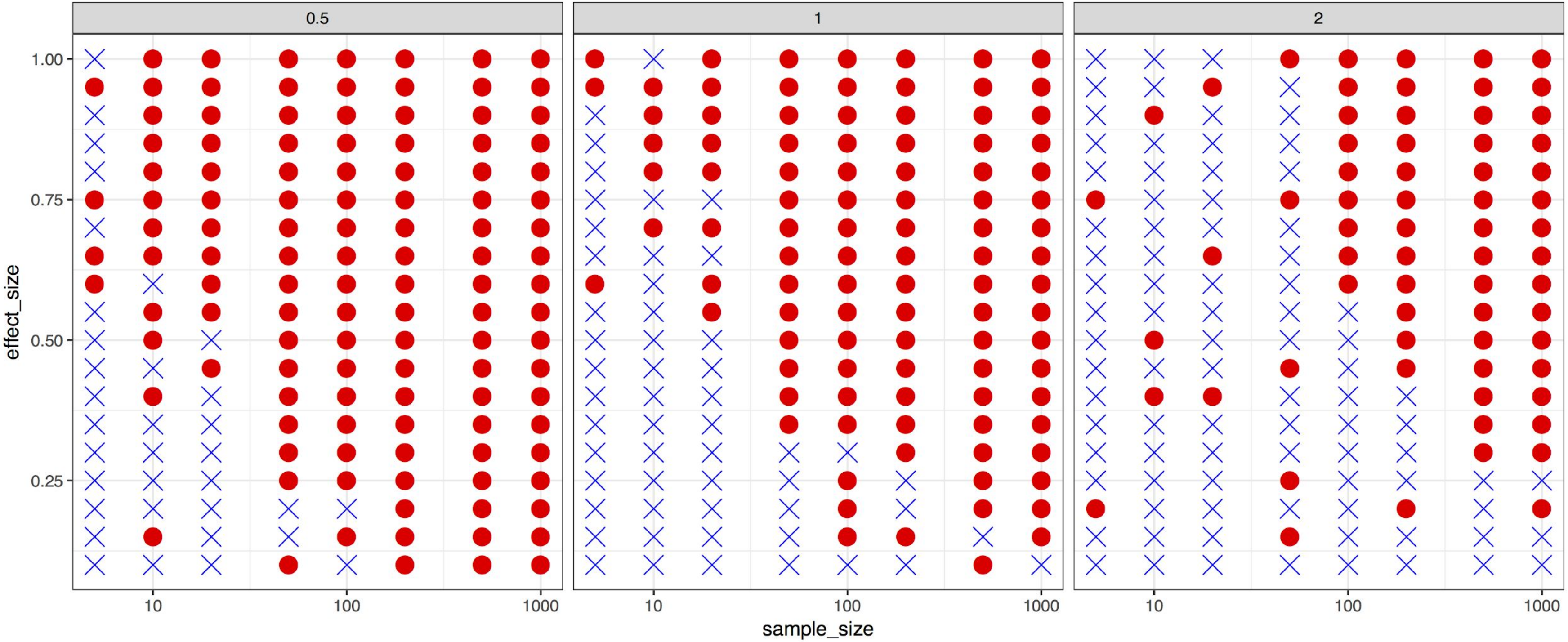
P-value depends on multiple factors

- P-values are dependent on: sample_size (effect_size = 0.25, std_deviation = 1)



P-value depends on multiple factors

- P-values are dependent on: sample_size, effect_size, within-group variance



P-value – Significant or not?

This list is culled from peer-reviewed journal articles in which:

a) the authors set themselves the threshold of 0.05 for significance,

b) failed to achieve that threshold value for p and

c) described it in such a way as to make it seem more interesting.

(barely) not statistically significant (p=0.052)

a barely detectable statistically significant difference (p=0.073)

a borderline significant trend (p=0.09)

a certain trend toward significance (p=0.08)

a clear tendency to significance (p=0.052)

a clear trend (p<0.09)

a clear, strong trend (p=0.09)

a considerable trend toward significance (p=0.069)

a decreasing trend (p=0.09)

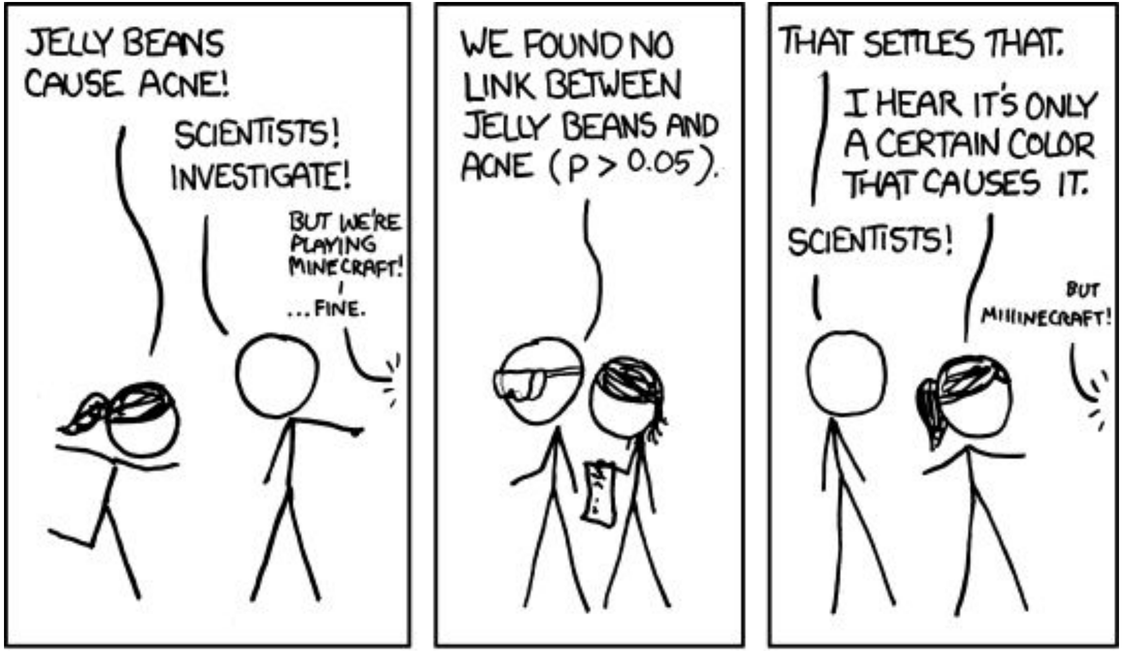
a definite trend (p=0.08)

a distinct trend toward significance (p=0.07)

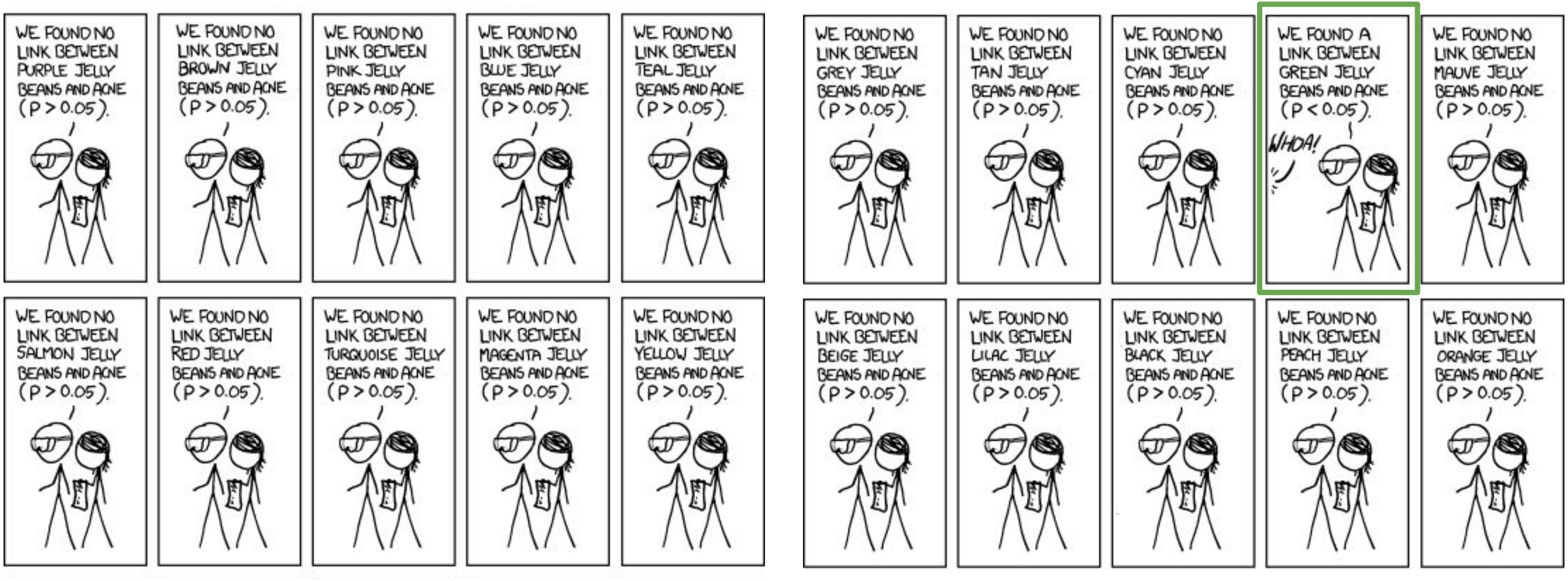
a favorable trend (p=0.09)

https://mchankins.wordpress.com/2013/04/21/still-not-significant-2/

Multiple hypothesis testing

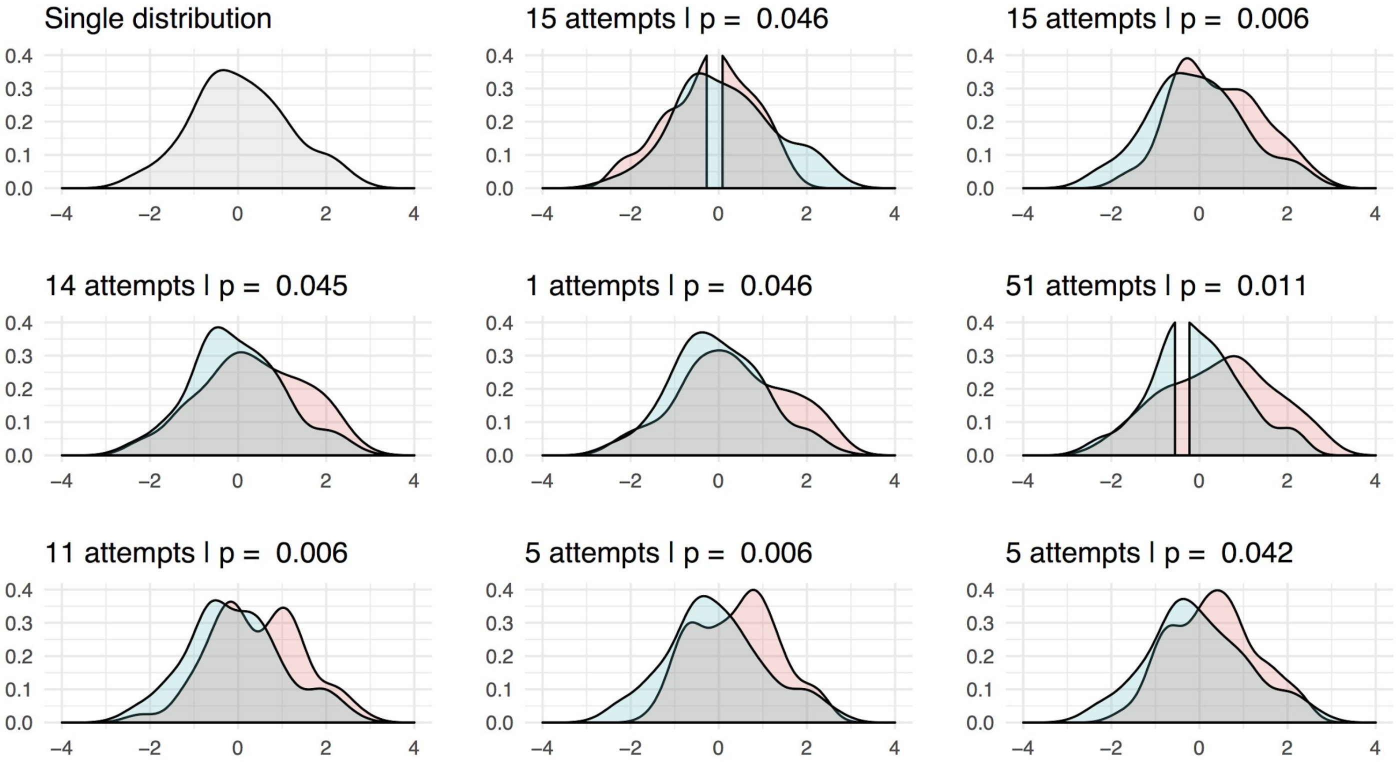

JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!

Multiple hypothesis testing



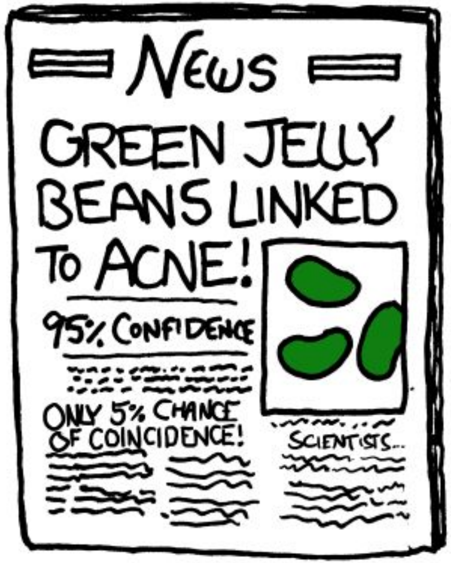
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).

Multiple hypothesis testing



“When a measure become a target, it ceases to be a good measure“ – Goodhart’s Law

Multiple hypothesis testing


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

The more inferences are made, the more likely erroneous inferences are to occur.

Let α be the Type 1 error rate for a statistical test.

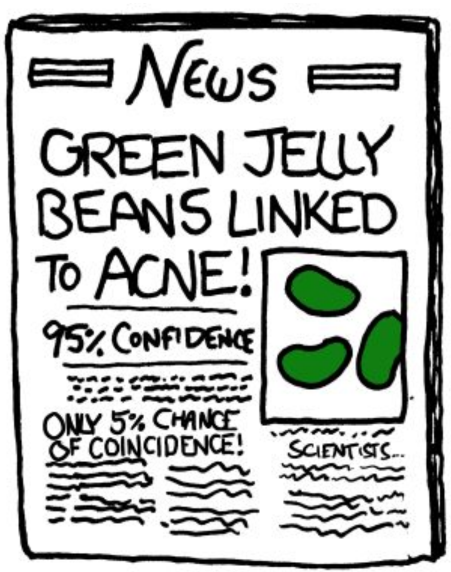
If the test is performed n times, what is the experimental-wise error rate α'? (Same as: What is the probability of obtaining at least 1 FP?)

$\alpha' = 1 - (1 - \alpha)^n$ (Check for $\alpha = 0.05$ & $n = 5$.)

The result may not be that significant even if its p-value < α.

To solve this problem, the nominal p-value need to be corrected/adjusted.

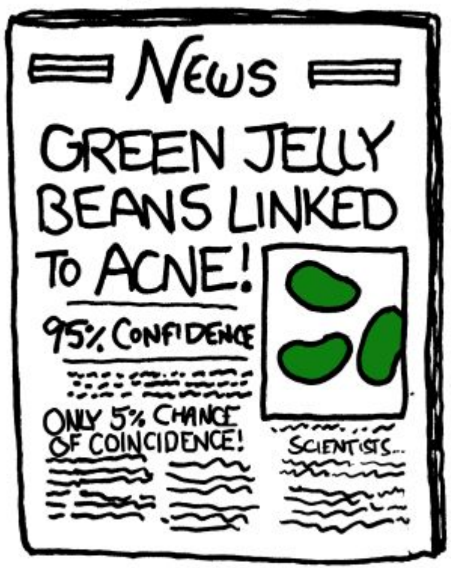
Correcting for multiple hypothesis testing


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Controlling for **Family-wise Error Rate**
(FWER: the probability of at least 1 FP):

- Bonferroni correction:
  - $p'_i = p_i * n$ (permutation test)
- Permutation test:
  - Permute the data K times, each time calculate minimum p-value
  - $p'_i = \#\{min\_pvalue < p_i\} / K$

Correcting for multiple hypothesis testing


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Controlling for **False Discovery Rate**
(FDR: proportion of FP among all significant hypotheses):

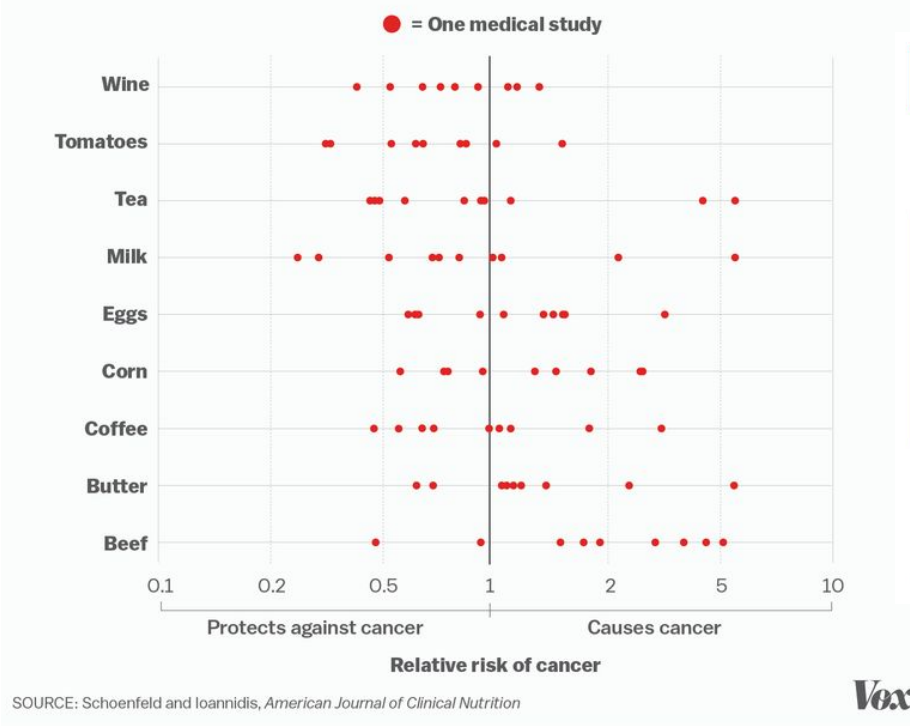
- Benjamini-Hochberg correction:
  - $p'_i = p_i * (n / i)$

Multiple hypothesis testing

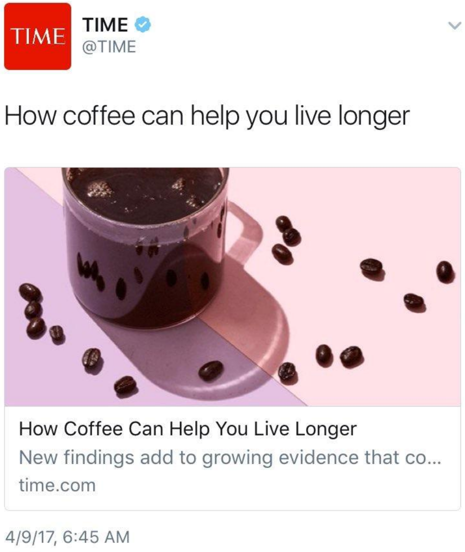
- FWER = Pr(#FP ≥ 1) = 1 – $(1 - \alpha)^n$. (Check for α = 0.05 & n = 5.)
- False discovery rate (FDR) = E[#FP / #Discoveries]
- Suppose 550 out of 10,000 genes are found to have different expression levels between disease and control samples at $p < 0.05$.
  - If p-value is chosen to control FWER, what is the #FP?
  - If p-value is chosen to control FDR, what is the #FP?

Multiple hypothesis testing

Publication bias (studies with nonsignificant results have lower publication rates)



SOURCE: Schoenfeld and Ioannidis, *American Journal of Clinical Nutrition*

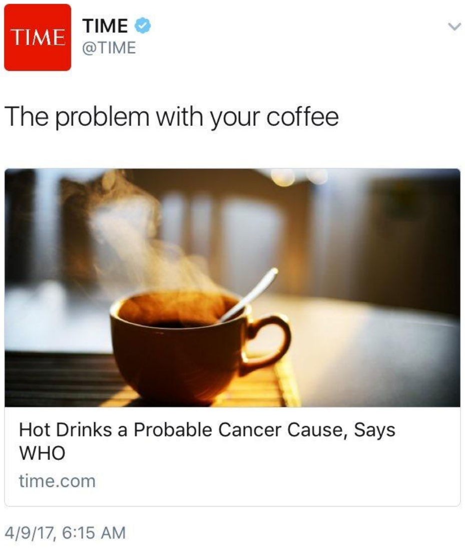


TIME
@TIME

How coffee can help you live longer

How Coffee Can Help You Live Longer
New findings add to growing evidence that co...
time.com

4/9/17, 6:45 AM



TIME
@TIME

The problem with your coffee

Hot Drinks a Probable Cancer Cause, Says WHO
time.com

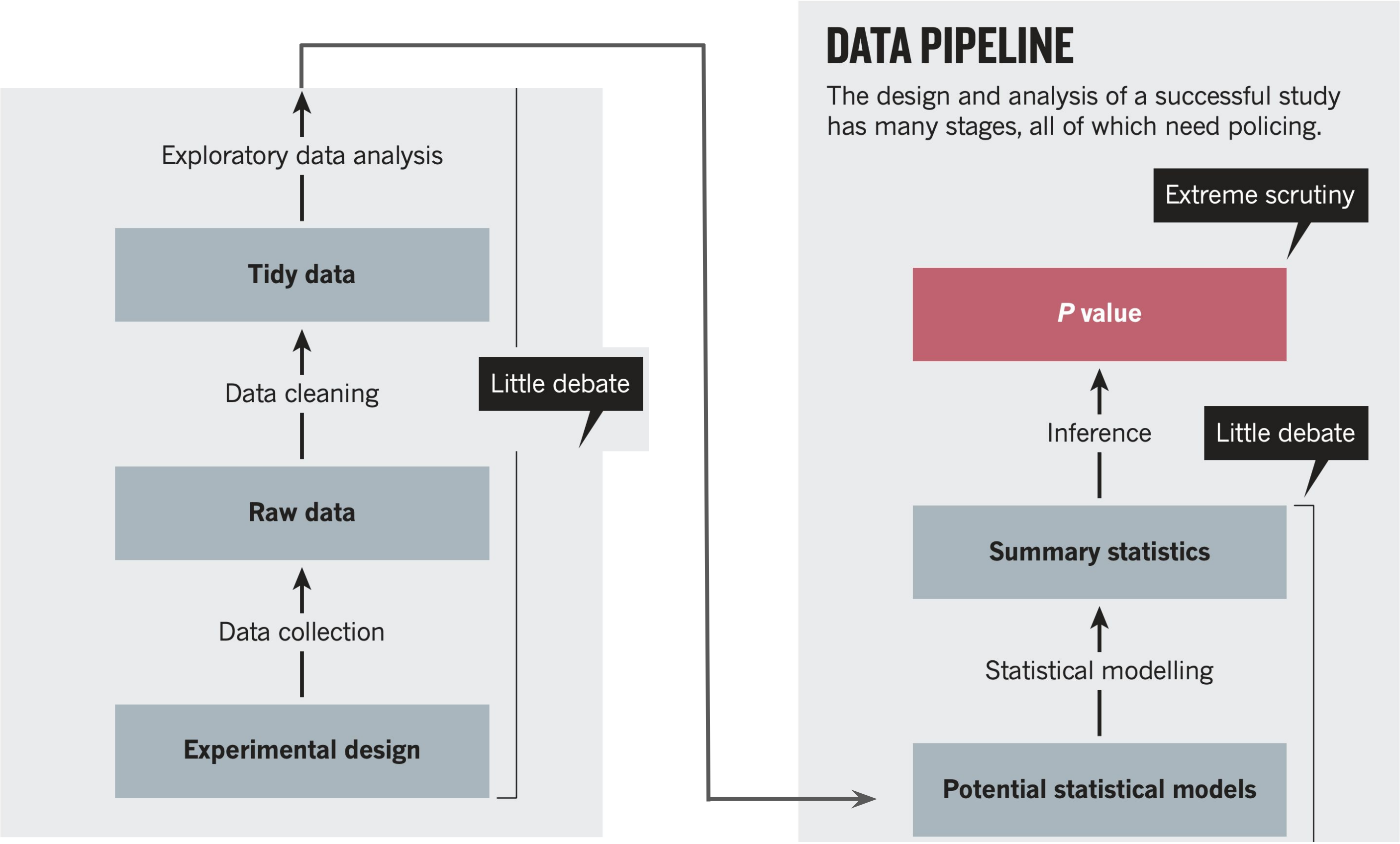
4/9/17, 6:15 AM

Questionable research practices

- Exclusively using p-values to determine the relevance and sanity of the results of a statistical test.
- Analyzing the data until the desired results are found.
- Collecting more data to reach smaller p-values.
- Trying many hypothesis until one of them gives a low p-value, and reporting just that final result.

WHEN YOU SEE A CLAIM THAT A COMMON DRUG OR VITAMIN "KILLS CANCER CELLS IN A PETRI DISH,"
KEEP IN MIND:
SO DOES A HANDGUN.

P-values are just the tip of the iceberg!



JT Leek, RD Peng http://www.nature.com/news/statistics-p-values-are-just-the-tip-of-the-iceberg-1.17412