Session 10: Hypothesis Testing, *continued*

Stats 60/Psych 10
Ismael Lemhadri
Summer 2020

Last time

- Hypothesis testing as a 6-step process
- One-sided and two-sided tests

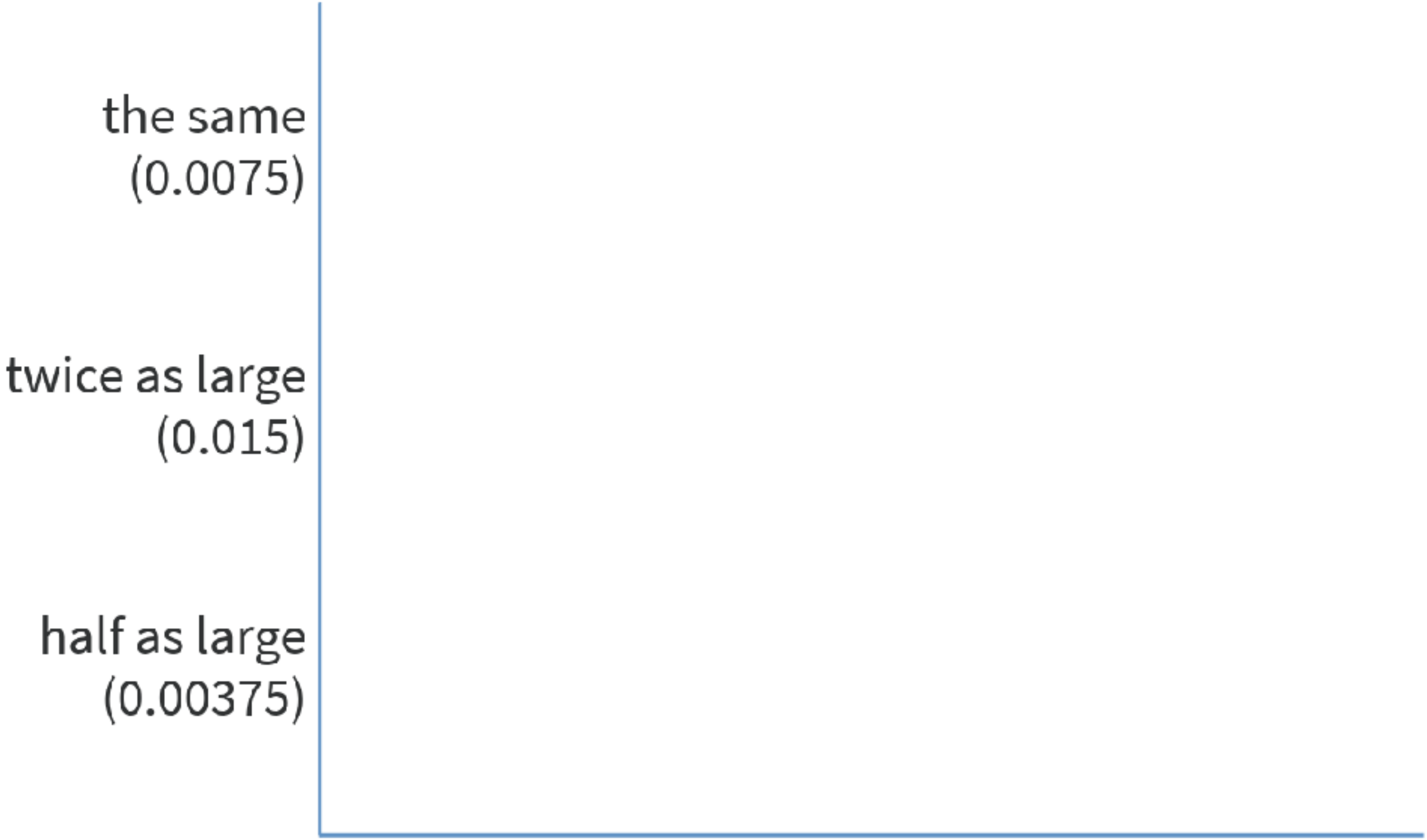
Last time

- Hypothesis testing as a 6-step process
- One-sided and two-sided tests

This time

- Assessing statistical significance
- The curse of multiple testing

If we reran the test as a two-tailed (non-directional) test, the p-value would be:



the same (0.0075)

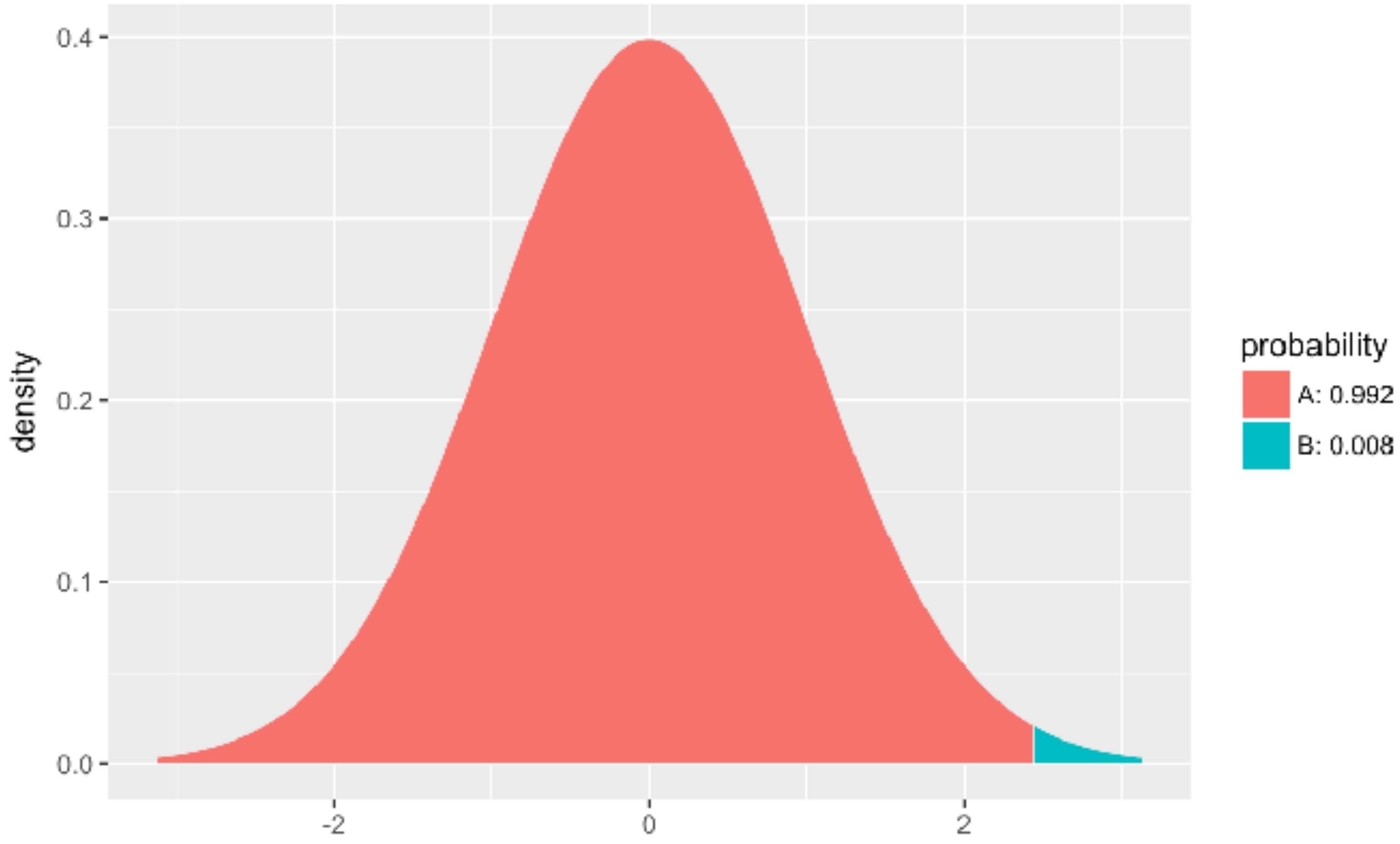
twice as large (0.015)

half as large (0.00375)

Start the presentation to see live content. Still no live content? Install the app or get help at PollEv.com/app

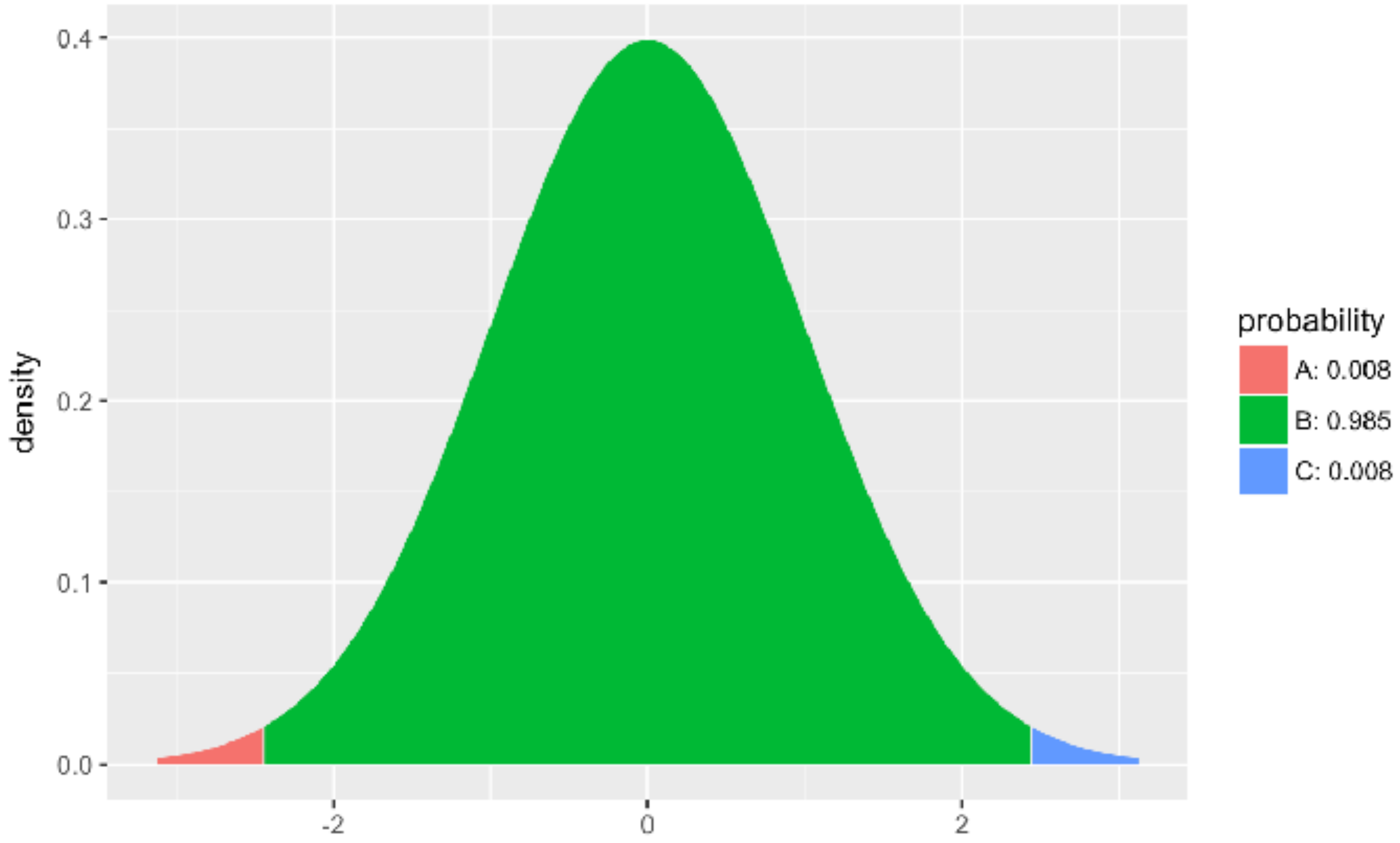
One-tailed vs two-tailed tests

- Directional test:
  - p-value = 1 - $p(t_{\text{observed}} \geq t_{248})$



One-tailed vs two-tailed tests

- Two-tailed (non-directional test)
  - p-value = 1 - p($t_{observed} \geq t_{248}$) + p($t_{observed} \leq t_{248}$)



Two-tailed results

```
ttestResult = t.test(BMI~PhysActive,data=NHANES_sample,var.equal=TRUE,
    alternative='two.sided')

   Two Sample t-test

data:  BMI by PhysActive
t = 2.4452, df = 248, p-value = 0.01517
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
 0.4329999 4.0193201
sample estimates:
mean of x mean of y
 29.63752  27.41136
```

p-value is twice as large for two-tailed test versus one-tailed test: data are less surprising!

Step 6: Assess the “statistical significance” of the result

- What does “statistical significance” mean?
- How much evidence against the null hypothesis do we require before rejecting it?

Sir Ronald Fisher

The (in)famous p<0.05

- “If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at .05”
- “it is convenient to draw the line at about the level at which we can say: Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials”

“the single most important figure in 20th century statistics” - Efron

p<0.05 was never meant to be a fixed rule

- Fisher:
    - “no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas”
- It probably became a ritual because of the difficulty in computing exact p-values in early days
    - All of the charts had entry for .05

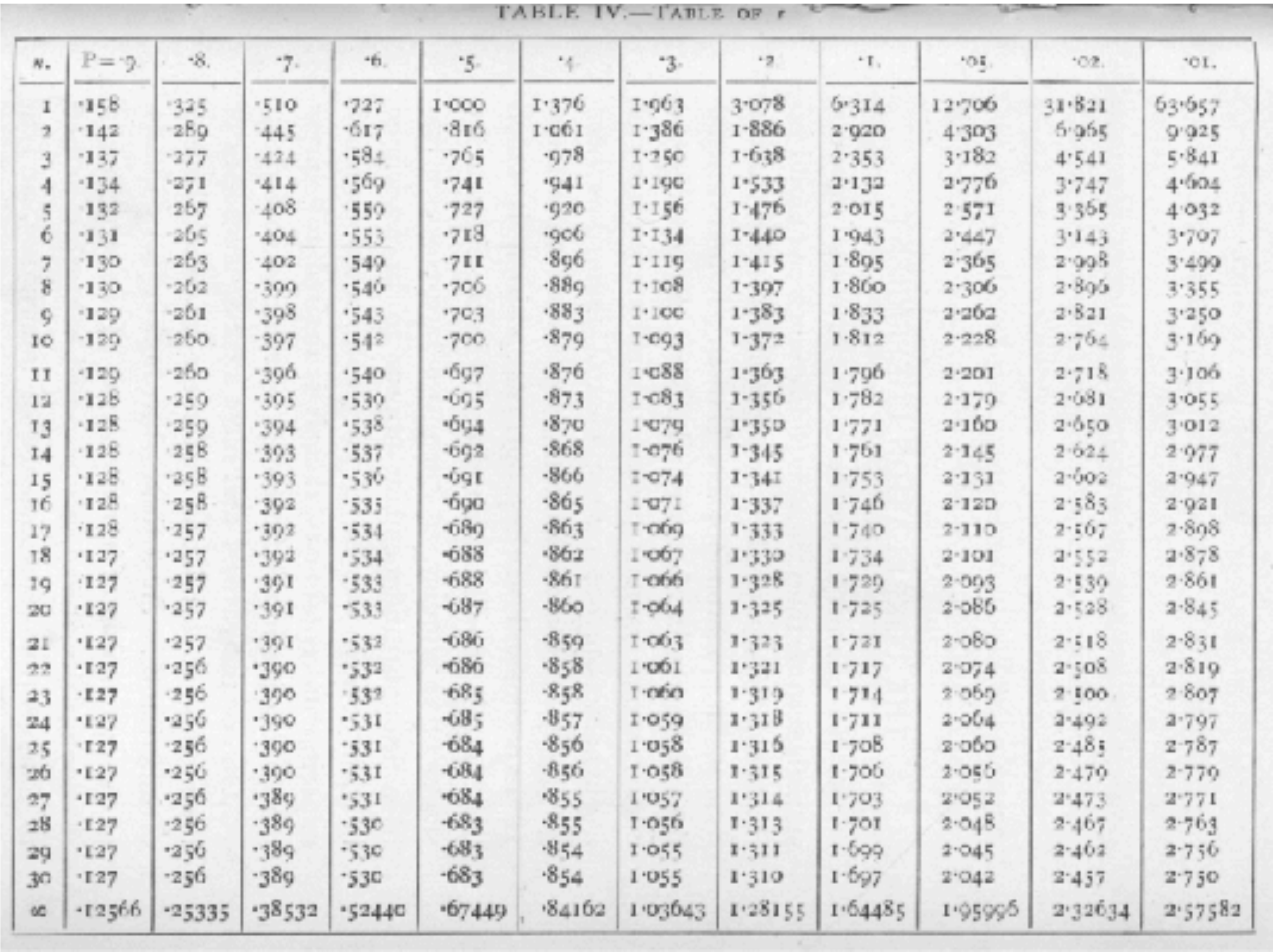
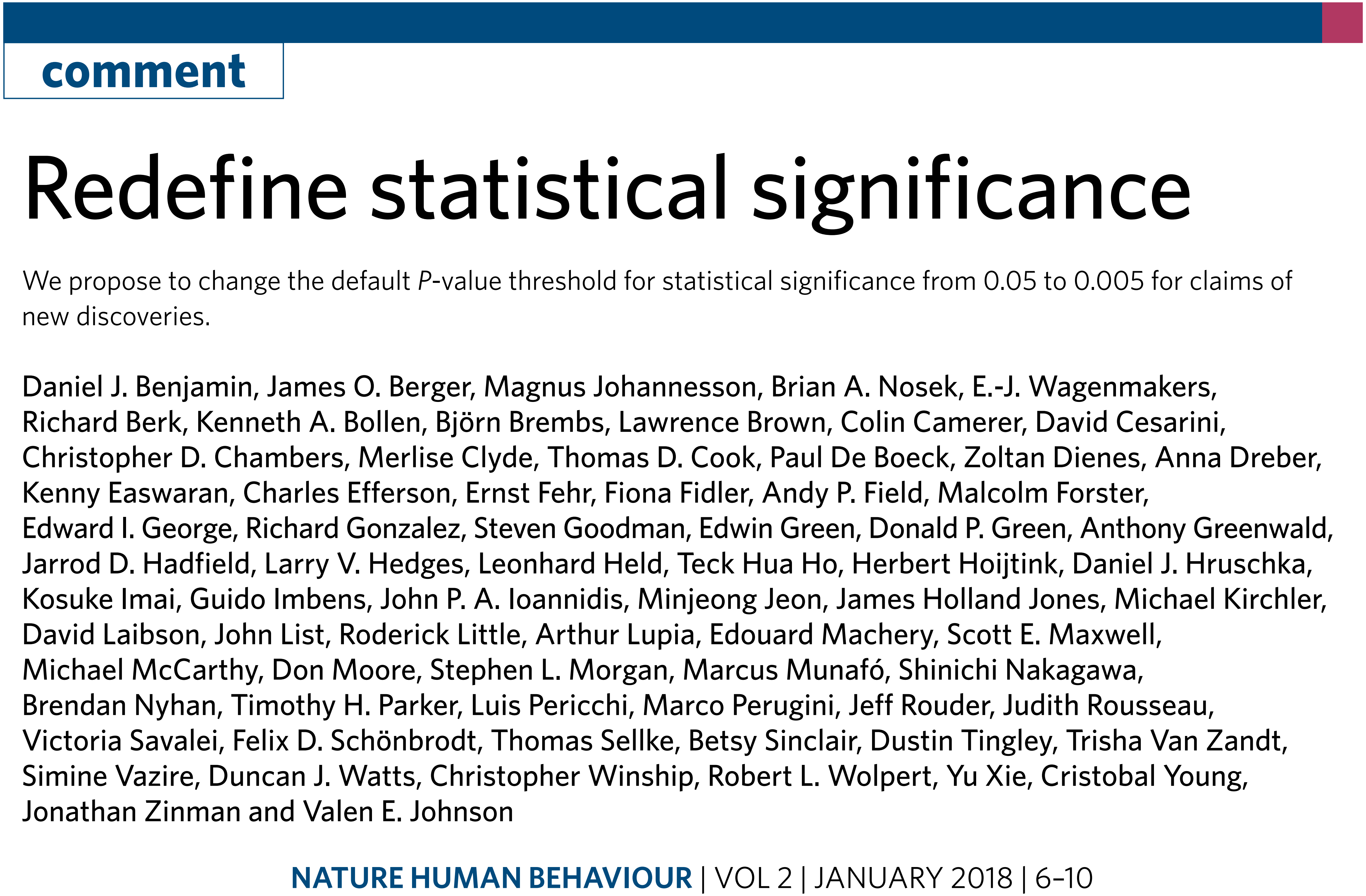


TABLE IV.—TABLE OF *t*

n.	P = ·9.	·8.	·7.	·6.	·5.	·4.	·3.	·2.	·1.	·05.	·02.	·01.
1	·158	·325	·510	·727	1·000	1·376	1·963	3·078	6·314	12·706	31·821	63·657
2	·142	·289	·445	·617	·816	1·061	1·386	1·886	2·920	4·303	6·965	9·925
3	·137	·277	·424	·584	·765	·978	1·250	1·638	2·353	3·182	4·541	5·841
4	·134	·271	·414	·569	·741	·941	1·190	1·533	2·132	2·776	3·747	4·604
5	·132	·267	·408	·559	·727	·920	1·156	1·476	2·015	2·571	3·365	4·032
6	·131	·265	·404	·553	·718	·906	1·134	1·440	1·943	2·447	3·143	3·707
7	·130	·263	·402	·549	·711	·896	1·119	1·415	1·895	2·365	2·998	3·499
8	·130	·262	·399	·546	·706	·889	1·108	1·397	1·860	2·306	2·896	3·355
9	·129	·261	·398	·543	·703	·883	1·100	1·383	1·833	2·262	2·821	3·250
10	·129	·260	·397	·542	·700	·879	1·093	1·372	1·812	2·228	2·764	3·169
11	·129	·260	·396	·540	·697	·876	1·088	1·363	1·796	2·201	2·718	3·106
12	·128	·259	·395	·539	·695	·873	1·083	1·356	1·782	2·179	2·681	3·055
13	·128	·259	·394	·538	·694	·870	1·079	1·350	1·771	2·160	2·650	3·012
14	·128	·258	·393	·537	·692	·868	1·076	1·345	1·761	2·145	2·624	2·977
15	·128	·258	·393	·536	·691	·866	1·074	1·341	1·753	2·131	2·602	2·947
16	·128	·258	·392	·535	·690	·865	1·071	1·337	1·746	2·120	2·583	2·921
17	·128	·257	·392	·534	·689	·863	1·069	1·333	1·740	2·110	2·567	2·898
18	·127	·257	·392	·534	·688	·862	1·067	1·330	1·734	2·101	2·552	2·878
19	·127	·257	·391	·533	·688	·861	1·066	1·328	1·729	2·093	2·539	2·861
20	·127	·257	·391	·533	·687	·860	1·064	1·325	1·725	2·086	2·528	2·845
21	·127	·257	·391	·532	·686	·859	1·063	1·323	1·721	2·080	2·518	2·831
22	·127	·256	·390	·532	·686	·858	1·061	1·321	1·717	2·074	2·508	2·819
23	·127	·256	·390	·532	·685	·858	1·060	1·319	1·714	2·069	2·500	2·807
24	·127	·256	·390	·531	·685	·857	1·059	1·318	1·711	2·064	2·492	2·797
25	·127	·256	·390	·531	·684	·856	1·058	1·316	1·708	2·060	2·485	2·787
26	·127	·256	·390	·531	·684	·856	1·058	1·315	1·706	2·056	2·479	2·779
27	·127	·256	·389	·531	·684	·855	1·057	1·314	1·703	2·052	2·473	2·771
28	·127	·256	·389	·530	·683	·855	1·056	1·313	1·701	2·048	2·467	2·763
29	·127	·256	·389	·530	·683	·854	1·055	1·311	1·699	2·045	2·462	2·756
30	·127	·256	·389	·530	·683	·854	1·055	1·310	1·697	2·042	2·457	2·750
∞	·12566	·25335	·38532	·52440	·67449	·84162	1·03643	1·28155	1·64485	1·95996	2·32634	2·57582

Fisher (1925)

Arguments against p<0.05



comment

Redefine statistical significance

We propose to change the default *P*-value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries.

Daniel J. Benjamin, James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, Daniel J. Hruschka, Kosuke Imai, Guido Imbens, John P. A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman and Valen E. Johnson

NATURE HUMAN BEHAVIOUR | VOL 2 | JANUARY 2018 | 6–10

Why is 0.05 problematic?

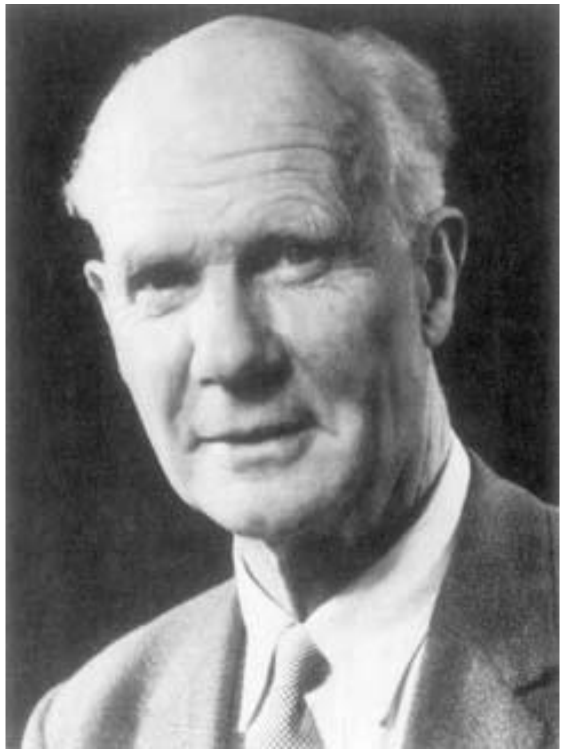
- $p<0.05$ indicates relatively weak evidence against the null
  - We will return to this later…

Statistical inference as decision making: Neyman/Pearson

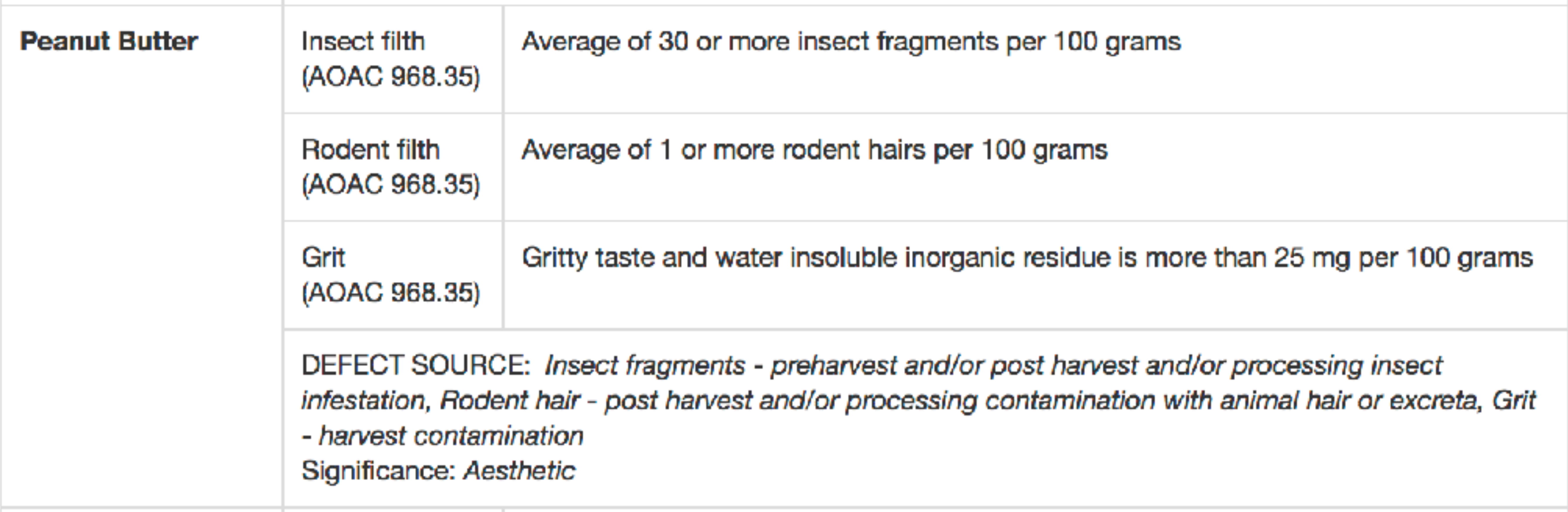
- “no test based upon a theory of probability can by itself provide any valuable evidence of the truth or falsehood of a hypothesis. But we may look at the purpose of tests from another viewpoint. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not often be wrong”
- We don’t know which specific decisions are right or wrong, but if we follow the rules, we know how often wrong decisions will occur

Jerzy Neyman

Egon Pearson



Example: statistical quality control



Peanut Butter	Insect filth (AOAC 968.35)	Average of 30 or more insect fragments per 100 grams
	Rodent filth (AOAC 968.35)	Average of 1 or more rodent hairs per 100 grams
	Grit (AOAC 968.35)	Gritty taste and water insoluble inorganic residue is more than 25 mg per 100 grams
	DEFECT SOURCE: *Insect fragments - preharvest and/or post harvest and/or processing insect infestation, Rodent hair - post harvest and/or processing contamination with animal hair or excreta, Grit - harvest contamination* Significance: *Aesthetic*	

https://www.fda.gov/Food/GuidanceRegulation/GuidanceDocumentsRegulatoryInformation/SanitationTransportation/ucm056174.htm

		Statistical decision	
		Reject H_0	Fail to Reject H_0
Reality	H_A is true	Correct (*hit*)	Type II error (*miss or false negative*)
	H_0 is true	Type I error (*false alarm* or *false positive*)	Correct (*correct rejection*)

P(Type I error) = $\boldsymbol{\alpha}$

The long-run probability of rejecting H_0 when it is true

		Statistical decision	
		Reject H_0	Fail to Reject H_0
Reality	H_A is true	Correct (*hit*)	Type II error (*miss* or *false negative*)
	H_0 is true	Type I error (*false alarm* or *false positive*)	Correct (*correct rejection*)

P(Type I error) = $\boldsymbol{\alpha}$

The long-run probability of rejecting H_0 when it is true

P(Type II error) = $\boldsymbol{\beta}$

The long-run probability of failing to rejecting H_0 when H_A is true

		Statistical decision	
		Reject H_0	Fail to Reject H_0
Reality	H_A is true	$1-\beta$ (statistical power)	β
	H_0 is true	α (false positive rate)	$1-\alpha$

alpha: How likely are we to reject H_0 when H_0 is true?

Statistical decision

		Reject H_0	Fail to Reject H_0
Reality	H_A is true	$1-\beta$ (statistical power)	β
	H_0 is true	α (false positive rate)	$1-\alpha$

alpha: How likely are we to reject H_0 when H_0 is true?

power: How likely are we to reject H_0 when H_A is true?

Breakout!

- Researchers generally set their false positive rate to 0.05, but their false negative rate (1-power) to 0.2
- Why might protecting from false positives be more important than protecting from false negatives?

Hypothesis testing demo

- In RStudio:

  -
    ```
    library(shiny)
    ```

  -
    ```
    runGitHub(“psych10/psych10”,
      subdir=“inst/hypothesis/“)
    ```

You run an experiment comparing means between two groups, and you find a significant difference (p=.01). Which of the following does this imply?

You have absolutely disproved the null hypothesis

You have found the probability of the null hypothesis being true

You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision

You have a reliable experimental finding in the sense that if the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions.

None of the above

Start the presentation to see live content. Still no live content? Install the app or get help at PollEv.com/app

What does a significant result mean?

- You run an experiment comparing means between two groups, and you find a significant difference (p=.01)
  - Does it mean that you have absolutely disproved the null hypothesis?
  -

What does a significant result mean?

- You run an experiment comparing means between two groups, and you find a significant difference (p=.01)
    - Does it mean that you have absolutely disproved the null hypothesis?
    - Does it mean that you have absolutely proved your experimental hypothesis?

What does a significant result mean?

- You run an experiment comparing means between two groups, and you find a significant difference (p=.01)
  - Does it mean that you have absolutely disproved the null hypothesis?
  - Does it mean that you have absolutely proved your experimental hypothesis?
    - No - statistics cannot prove or disprove hypotheses!
    - It provides relative evidence against the null

What does a significant result mean?

- Does it mean that you have found the probability of the null hypothesis being true?
- Does it mean that you can deduce the probability of the altnernative hypothesis being true?
    - No: The p-value is the probability of the data, not the probability of any hypothesis
        - p-value = $P(D|H_0)$
        - If we want to know $P(H_0|D)$, what do we need to use?
        - And what do we need to know in order to use it?

What does a significant result mean?

- Does it mean that you know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision?
  - Restate this: $P(H_0 \text{ is true}|p<\text{alpha})$?

What does a significant result mean?

- Does it mean that you know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision?
  - Restate this: $P(H_0 \text{ is true} | p < \text{alpha})$?
  - p-values are probabilities of data, not hypotheses!

NHST in a modern context

- Null hypothesis statistical testing can become very challenging in the context of modern science and big data
- Traditionally, researchers measured very few variables on each individual
- In modern science, we can often measure millions of variables per individual
    - Genomics
    - Brain imaging

A real-life example of hypothesis testing in action

- We know that schizophrenia has a strong genetic basis
    - About 80% of variation in schizophrenia is due to genetic differences
- Research has begun to look at which specific genes are involved
    - Look at many places in the genome where people differ in their genetic code (“polymorphisms”)
        - Usually about 1 million different locations
    - Test whether people with schizophrenia are more likely to have a different version of the genetic code at that location

The problem with multiple hypothesis tests

- Let’s say we did 1 million hypothesis tests at $p < 0.05$
    - # of expected errors if the null hypothesis is true
        - N * alpha = 1,000,000 * 0.05 = 50,000
- $p < 0.05$ is appropriate to control the error rate for a single test
- What we really want to control is the “familywise error rate” - the likelihood of at least one false positive across our entire “family” of tests
- With 1 million tests at $p < 0.05$, the familywise error rate will be ~1
    - Every study will have false positives

Controlling for multiple comparisons

- If all of the tests are independent, we can control this by dividing our alpha level by the number of tests
  - “Bonferroni correction”
  - For 1 million tests, this would be:
    - $p < 0.05/1{,}000{,}000$ (5e-08)
  - This ensures that we expect a false positive finding in only 1 out of every 20 studies

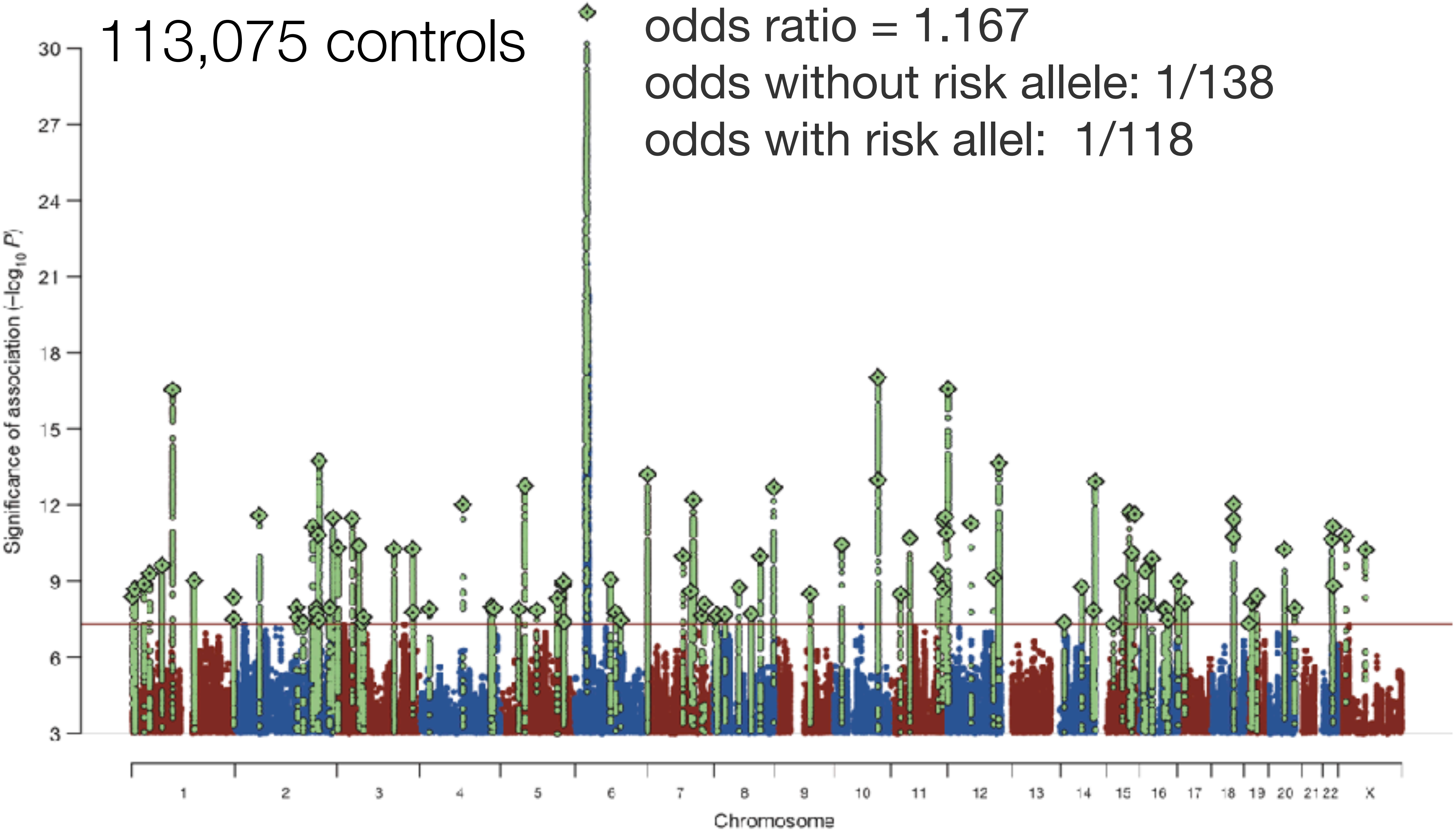
Simulating the effects of multiple testing

```
nTests=10000

uncAlpha=0.05
uncOutcome=replicate(nTests,
          sum(rnorm(nTests)<qnorm(uncAlpha)))

print(paste('uncorrected:',mean(uncOutcome>0)))
[1] "uncorrected: 1"

corAlpha=0.05/nTests
corOutcome=replicate(nTests,
           sum(rnorm(nTests)<qnorm(corAlpha)))

print(paste('corrected:',mean(corOutcome>0)))
[1] "corrected: 0.047"
```

“Manhattan plot” of genetic associations with schizophrenia

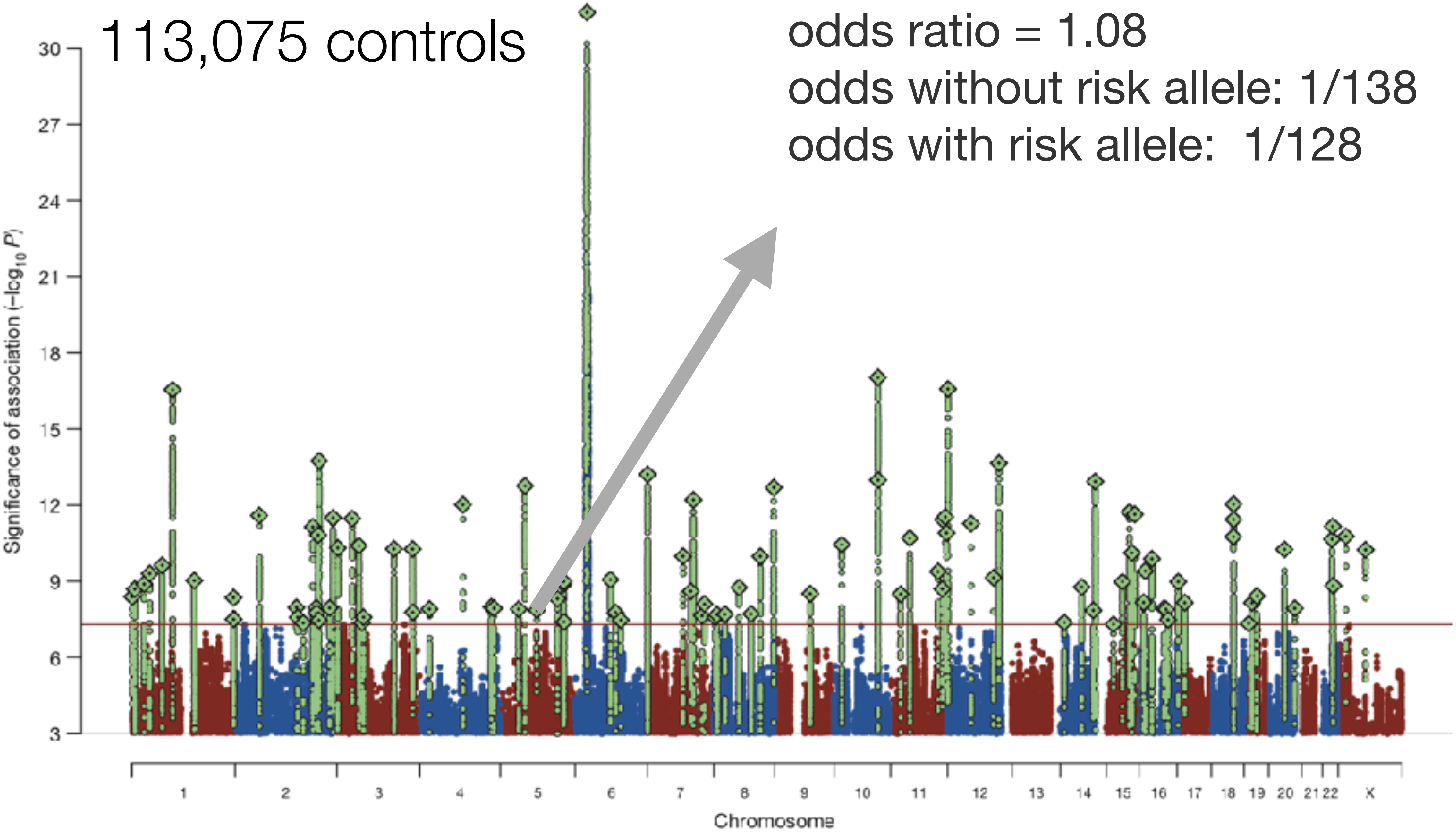
36,989 cases
113,075 controls

odds ratio = 1.167
odds without risk allele: 1/138
odds with risk allel: 1/118



PGC, 2014

“Manhattan plot” of genetic associations with schizophrenia



36,989 cases
113,075 controls

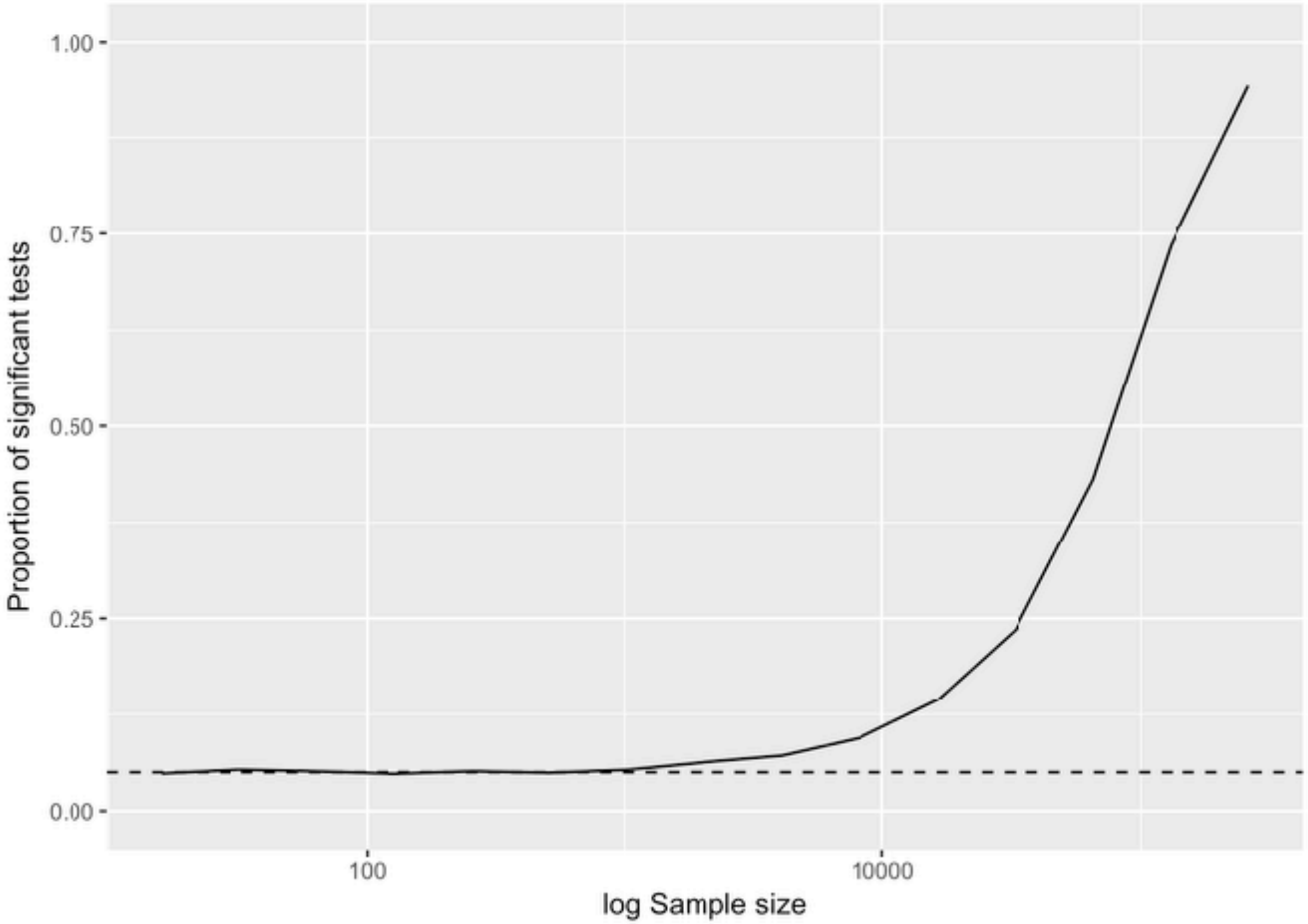
odds ratio = 1.08
odds without risk allele: 1/138
odds with risk allele: 1/128

PGC, 2014

Statistical significance and sample size

- Meehl's paradox
  - In many areas of science (such as physics), higher N provides more precise models
  - Using NHST, as N becomes large, everything becomes significant

True effect size = 0.01 SD



Recap

- We can use statistics to test hypotheses
- P-values provide us a measure of how surprising the data would be if there was truly no effect
    - They do not necessarily tell us how strong the effect is
- We can use either theoretical distributions or randomization to determine the distribution of our statistic under the null hypothesis
- When we perform multiple tests, we have to adjust our threshold to prevent inflation of false positive rates