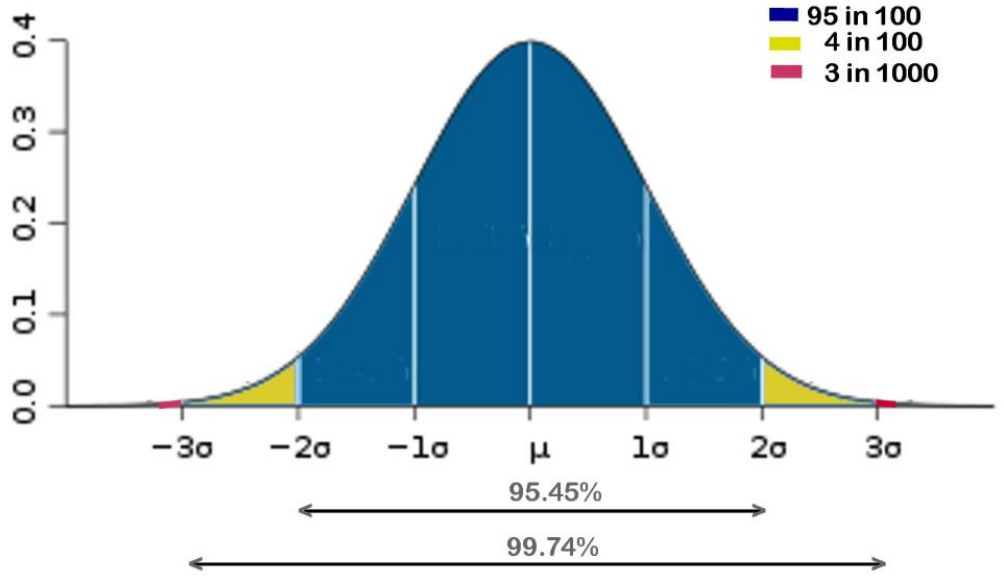
Hypothesis Testing

Seth D. Temple
Statistics Ph.D. candidate
UW Biostatistics Teaching Demo

Last class (param. vs. statistic)

- Measures of central tendency
  - Mean
  - Median
  - Mode
- Measure of spread
  - Standard deviation
  - Interquartile range



Learning objectives

- Frame a research question as statistical question
- Define Type 1 and Type 2 Error
- Discuss their importance in research study

Participation expectations

- Raise your hand to ask me for clarification
- Exercise 1
  1. Chatter in small group, or write response on notes (~30 sec.)
  2. Solicit responses for larger discussion

- Exercise 2
  - Think about your response privately (5 sec.)
  - Quick “Yes/No” poll after countdown

- Show respect; no judgment

???

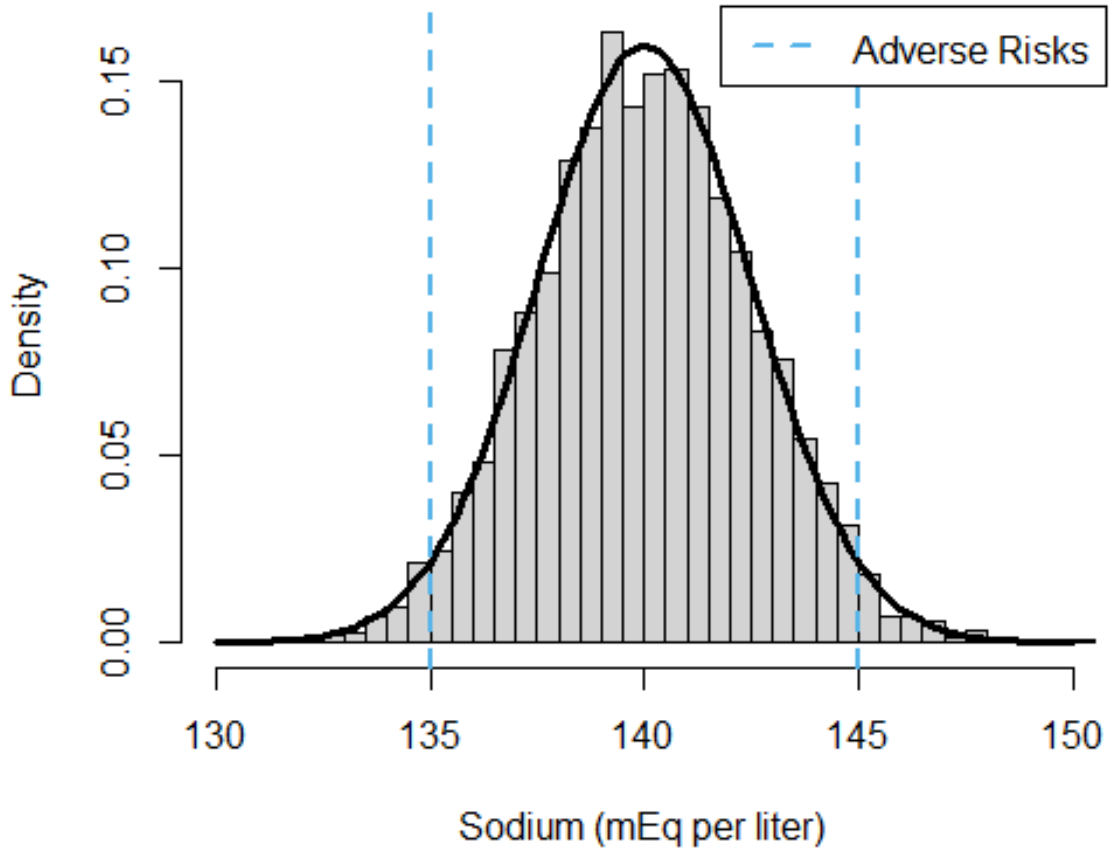
Warm-up viral tests

		Actual Values: Positive (1)	Actual Values: Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Example 1: Sodium homeostasis

- Sodium is important to biological processes
- Too little may lead to health consequences

- Lamotrigine helps treat bipolar disorder. It may affect sodium levels.



Research Question

- How does **lamotrigine** affect the levels of **sodium metabolite levels**?

Statistical Question

- Among a sample of people taking **lamotrigine**, do observed levels of **sodium** differ from typical value derived from prior medical study?

Example 1: Sodium homeostasis

Model for sodium level

- $\mu_0 = 140$ is “typical”, data looks like a bell curve, $\sigma = 2.5$
- Initial assumption is $\mu_D = \mu_0 = 140$.

Experiment

- Collect sodium data from people taking D.
- If average $\bar{X}_D$ is very different from 140, we conclude $\mu_0 \neq \mu_D$.

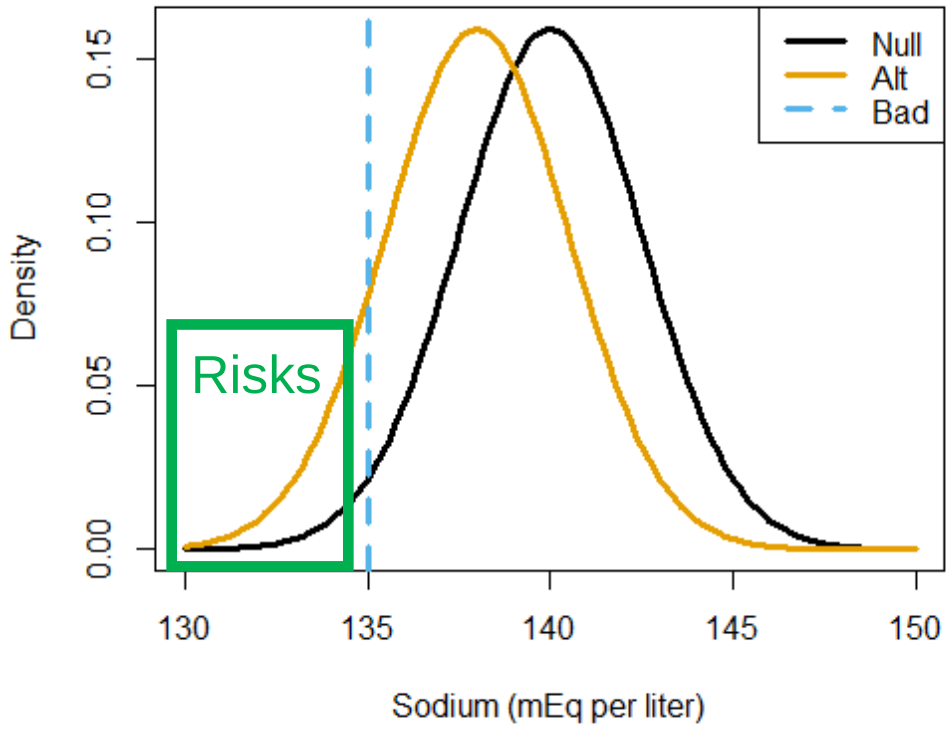
Why do we care about hypothesis tests?

Reward

- Want to prescribe patients of an effective treatment

Risk

- Don’t want to impose risks to lamotrigine takers



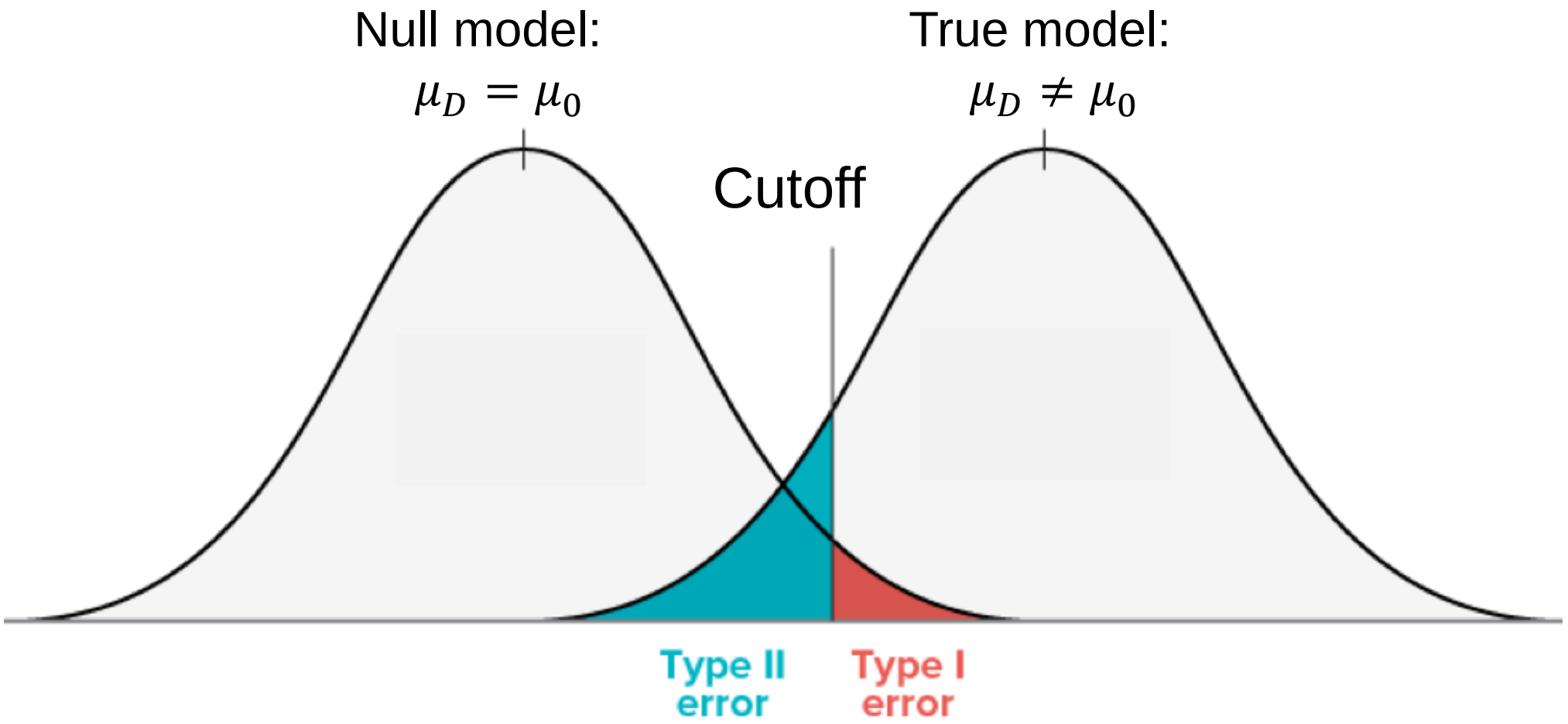
Type 1 Error. The probability we reject $\mu_D = 140$ though the model is true.

* This definition generalizes to other hypothesis tests.

Type 2 Error. The probability we fail to reject $\mu_D = 140$ given that $\mu_D = 138$.

* This definition generalizes to other hypothesis tests, say $\mu_D \neq 140$.

Visualizing Type I,II errors



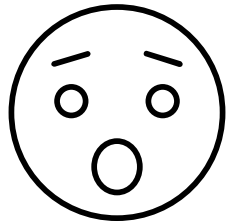
Type 1 error calculation

If I experiment again, probability I observe average sodium as extreme as $\bar{X}_D$ among 30 samples.

$$\bar{X}_D = 138.42$$

$$Z_D = \sqrt{30} \times \frac{(\bar{X}_D - 140)}{2.5} = -3.46$$

(p-value) Probability more extreme is 0.0005 !!!



Type 1 error simulations

- Assume model $\mu_0 = 140$, $\sigma = 2.5$, bell curve
- Run 1000 experiments
  1. Simulate 30 samples from model
  2. Compute Z_D statistic
  3. If absolute value of $Z_D > 1.96$, add 1 to running count

I found $Z_D > 1.96$ to occur $\frac{53}{1000} \approx 0.05$.

Discussion activity (30 sec.)

Study effect of lamotrigine on # **of mood swings**

- 50 participants in Drug D group
- 50 participants in Placebo P group

1. How would you phrase statistical question in this study?
2. How would you phrase the Type 1 error?

My answer key

1. Among sample of individuals diagnosed with bipolar disorder, do observed # of mood swings $\bar{X}_D - \bar{X}_P$ differ from typical difference $\mu_D - \mu_P$?

2. The probability that we reject initial assumption $\mu_D = \mu_P$ despite the true model being $\mu_D = \mu_P$.

Alternative Hypothesis Experiment

- Run an experiment 2000 times
  1. Simulate 30 samples from model Normal(μ_D, $\sigma = 2.5$)
  2. Compute $Z_D = \sqrt{30} \times (\bar{X}_D - 140)/2.5$
  3. If absolute value of $Z_D > 1.96$, add 1 to a running count

True Drug Effect	Type 2 Error	Power = 1 - Type 2
139.75	0.91	0.09
139.50	0.81	0.19
139.00	0.42	0.58
138.00	0.01	0.99

Sample Size Experiment

Same as before, except # of people varies

The true drug effect is 139

Sample Size	Type 2 Error
10	0.77
20	0.57
30	0.42
50	0.28

Confidence level experiment

Same as before, except controlling Type 1 error varies

The true drug effect is 139

Confidence	Type 2 Error
0.10	0.29
0.05	0.41
0.01	0.65
0.001	0.86

Why would we want to have lower confidence level if it decreases Type 2 error?

(30 seconds)

My answer key

Type II error depends on specifying tolerance for Type I error.

We may want to make few Type 1 errors.

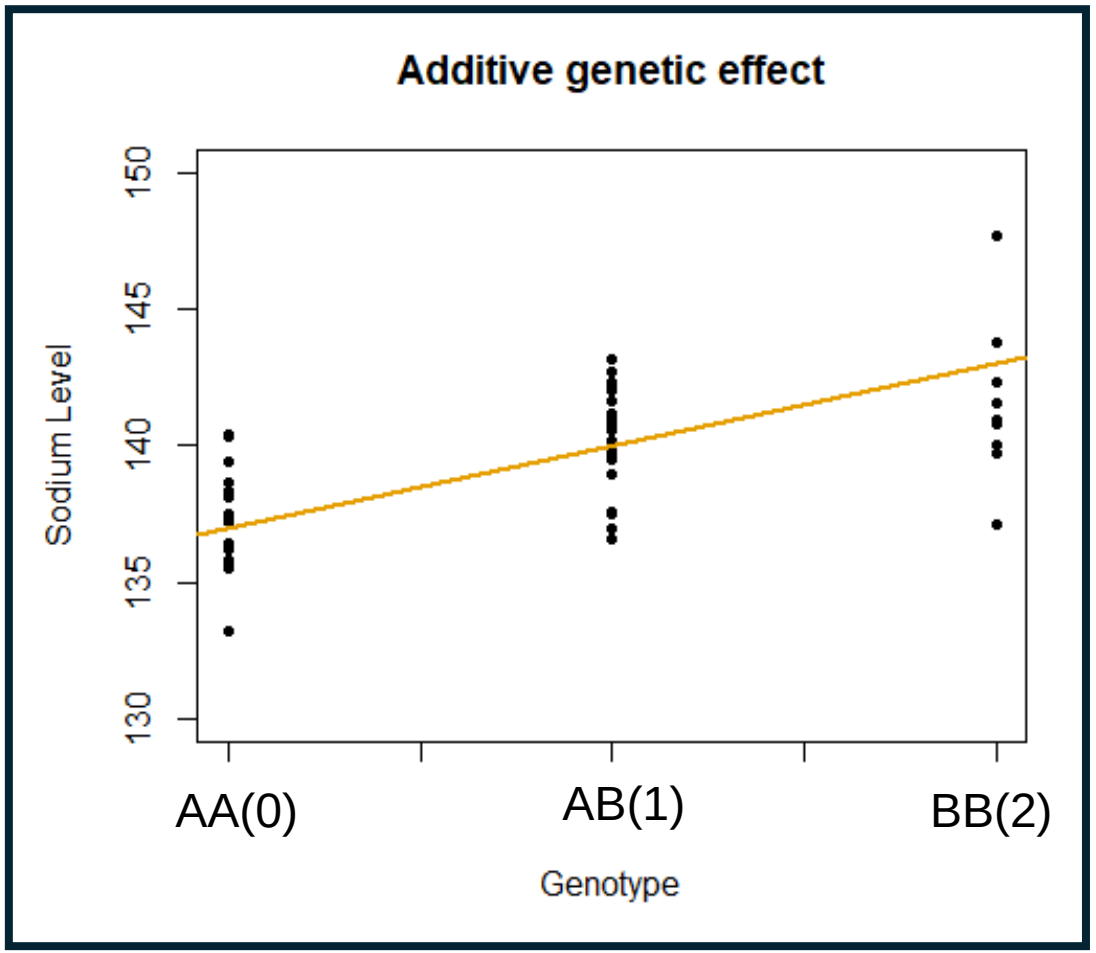
Beyond p-values

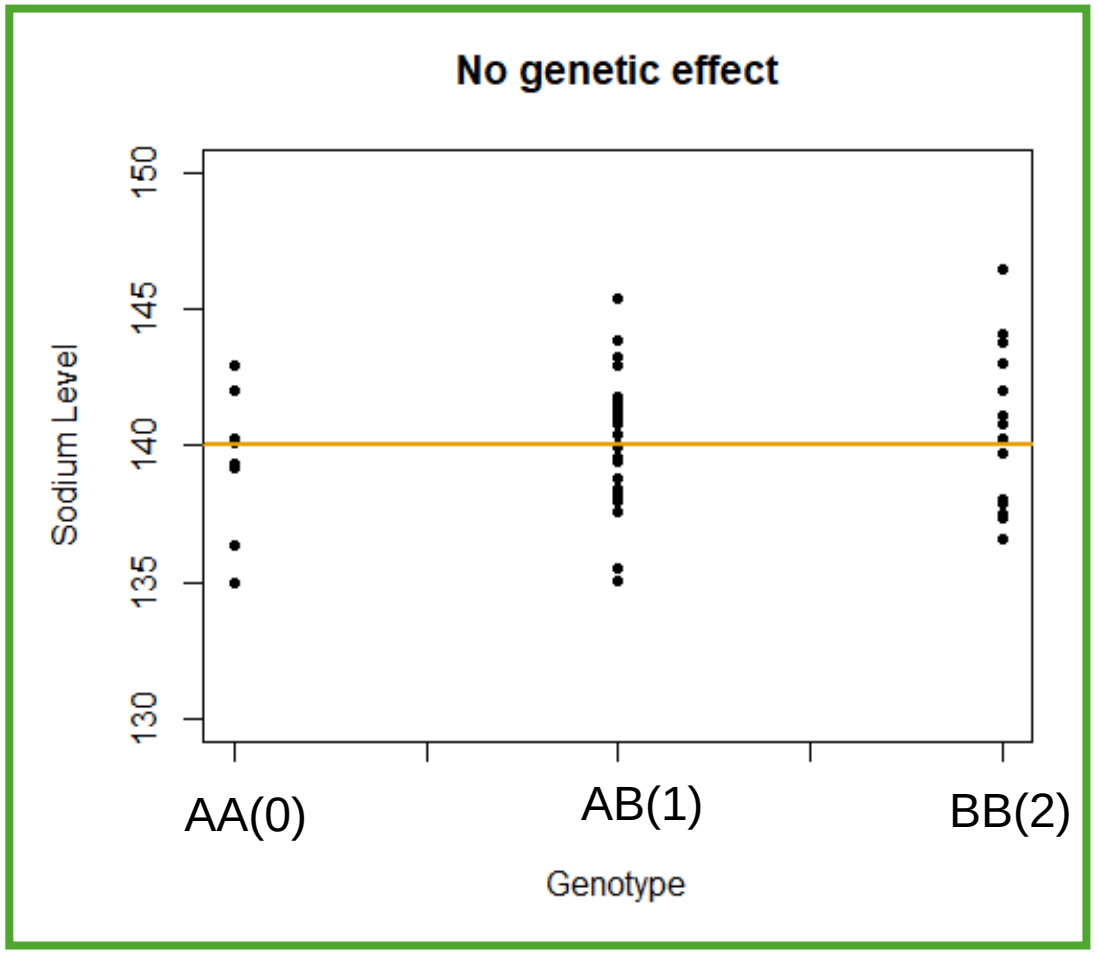
- Lamotrigine shows “significant” effect on mood!
  - How effective is significant? 😕
  - “Typical” is based on non-representative study. 😕
  - Prescription costs 😕
  - Studies had few participants ☹
  - Electrolytes are effective supplement ☺

Example 2: Genetic associations

- Genotypes AA (0), AB (1), BB (2)
- $(Y_1, \ldots, Y_n)$ sodium levels
- $(X_1, \ldots, X_n)$ genotypes
- ϵ_i random noise

$$Y_i = \underbrace{\beta_0}_{\text{Intercept}} + \underbrace{\beta_1}_{\text{Slope}} \times X_i + \epsilon_i$$


Additive genetic effect
Sodium Level
130
135
140
145
150
AA(0)
AB(1)
BB(2)
Genotype


No genetic effect
Sodium Level
130
135
140
145
150
AA(0)
AB(1)
BB(2)
Genotype

???

Studying one locus

Simulated no effect model:

$$Y = 140 + 0 \times X + \epsilon$$

Parameter	Estimate	p-value
Slope	-0.09	0.67

???

Studying one locus

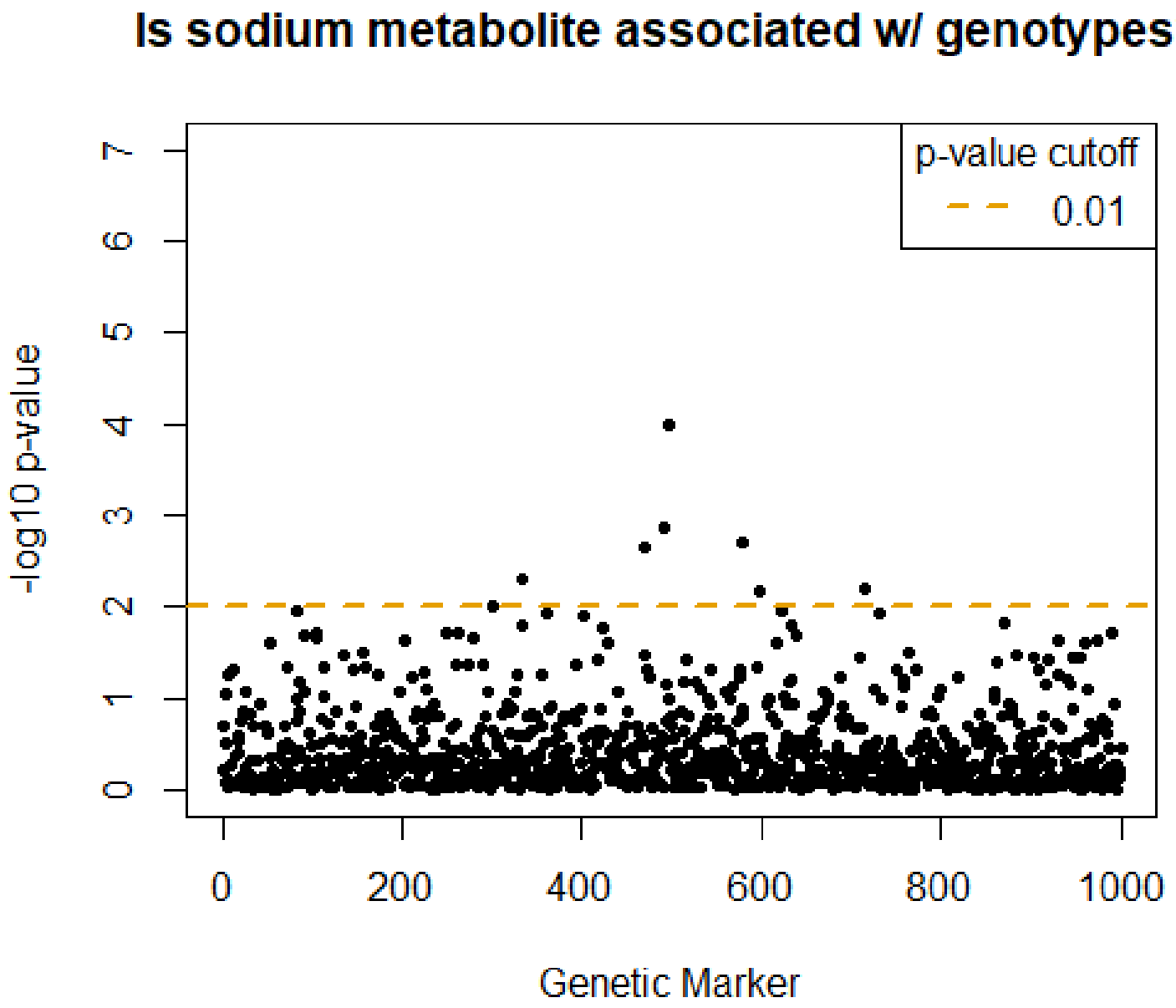
Simulated additive effect model:

$$Y = 137 + 3 \times X + \epsilon$$

Parameter	Estimate	p-value
Slope	2.91	2e-16

Simulate no genetic effect

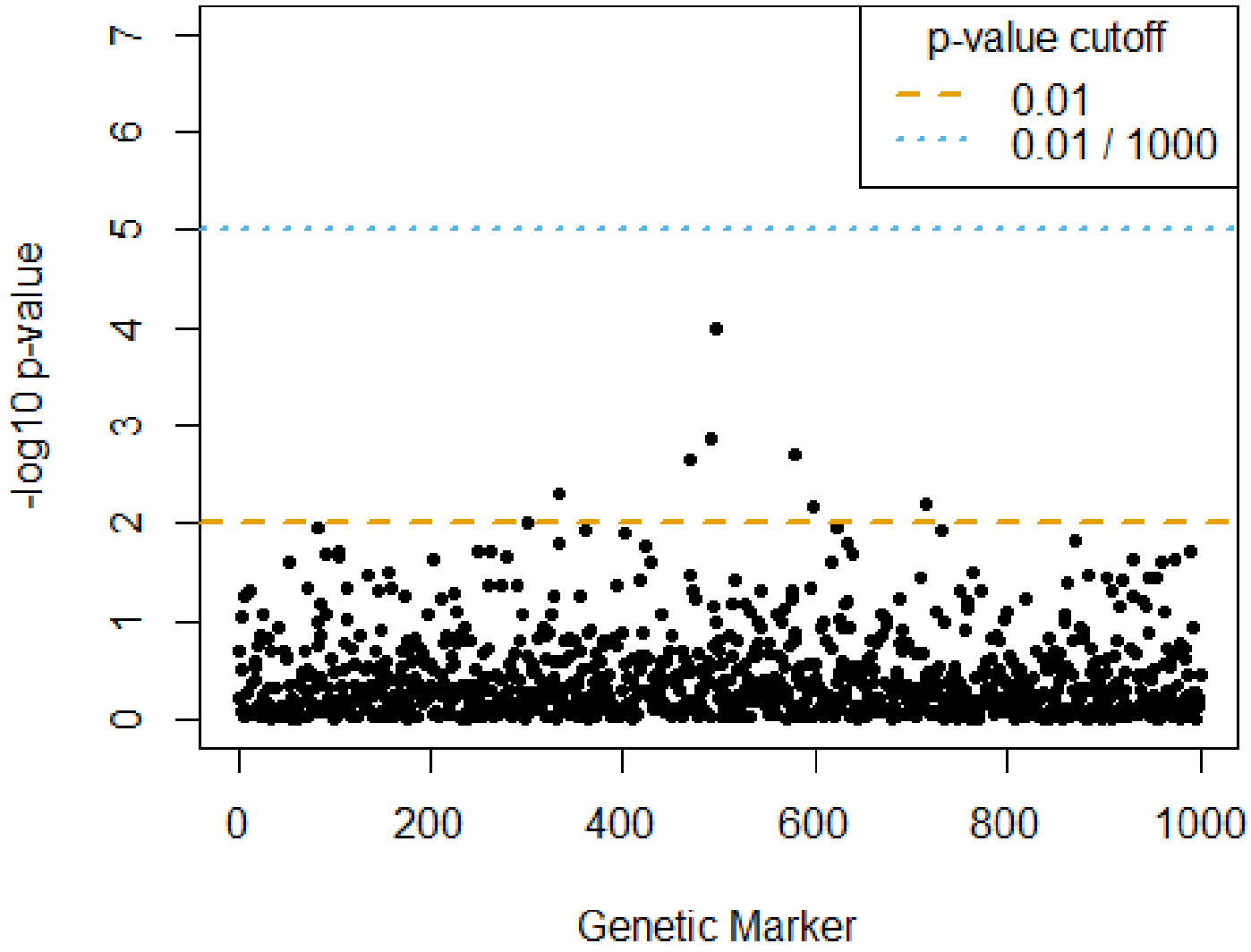
1. Draw 16, 17, 17 genotypes AA, AB, BB
2. Draw 50 errors $\epsilon \sim N(\mu = 0, \sigma = 2.5)$
3. Compute Y
   - Intercept $\beta_0 = 140$
   - Slope $\beta_1 = 0$
4. Compute p-value from average $\bar{Y}$


Is sodium metabolite associated w/ genotypes?
p-value cutoff
0.01
-log10 p-value
0
1
2
3
4
5
6
7
0
200
400
600
800
1000
Genetic Marker

Why do we observe ~ 10 / 1000 “significant” associations? (30 seconds)

- p-value threshold should be smaller
- Divide by # tests

Is sodium metabolite associated w/ genotypes?



Today's review

- Two examples of statistical analysis
  - Average sodium level (continuous)
  - Genetic association (linear relationship)

- **Type 1 error** is controlling for false positives
  - Testing too many things
- **Power** is identifying true negatives

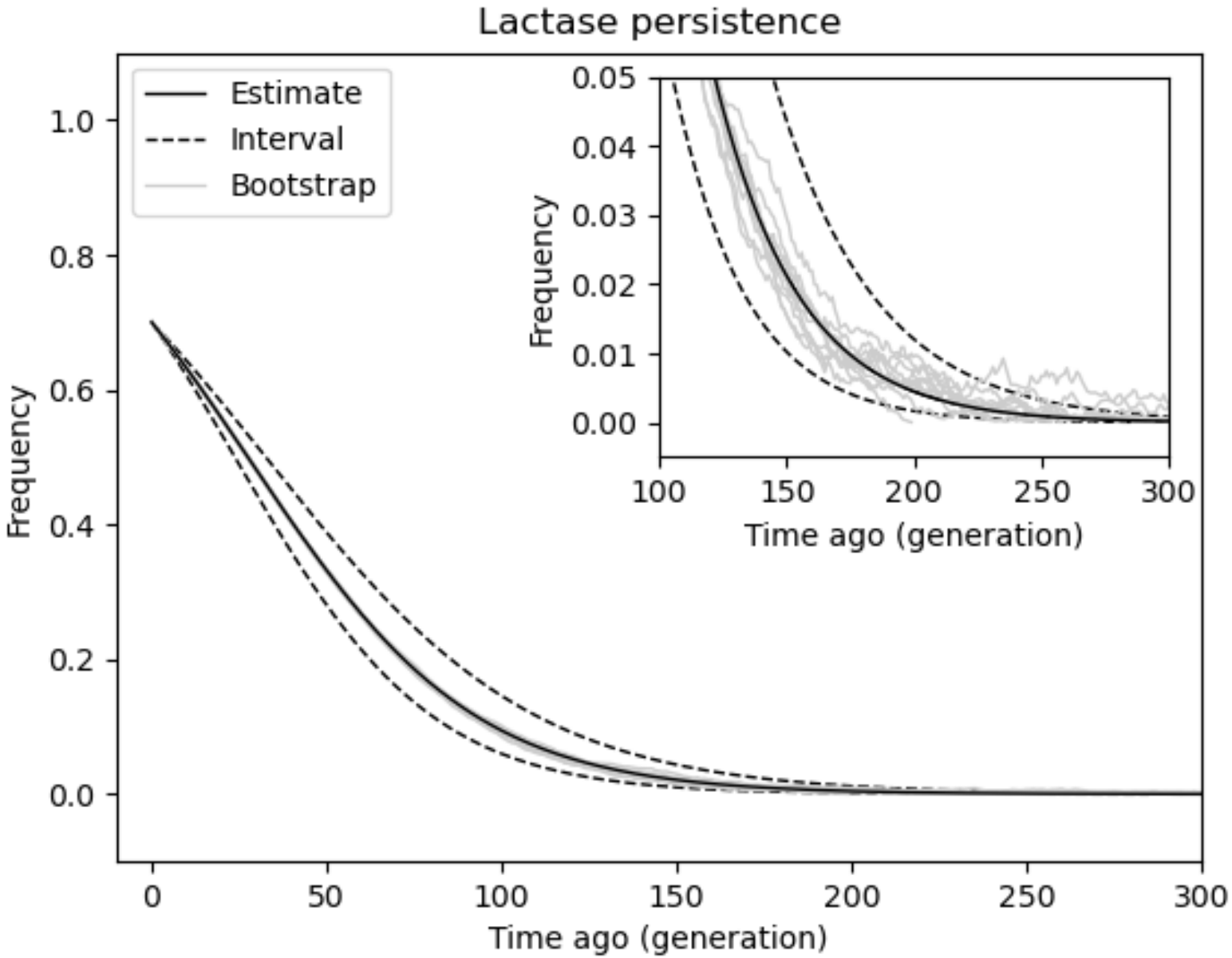
Next class: more types of hypothesis tests

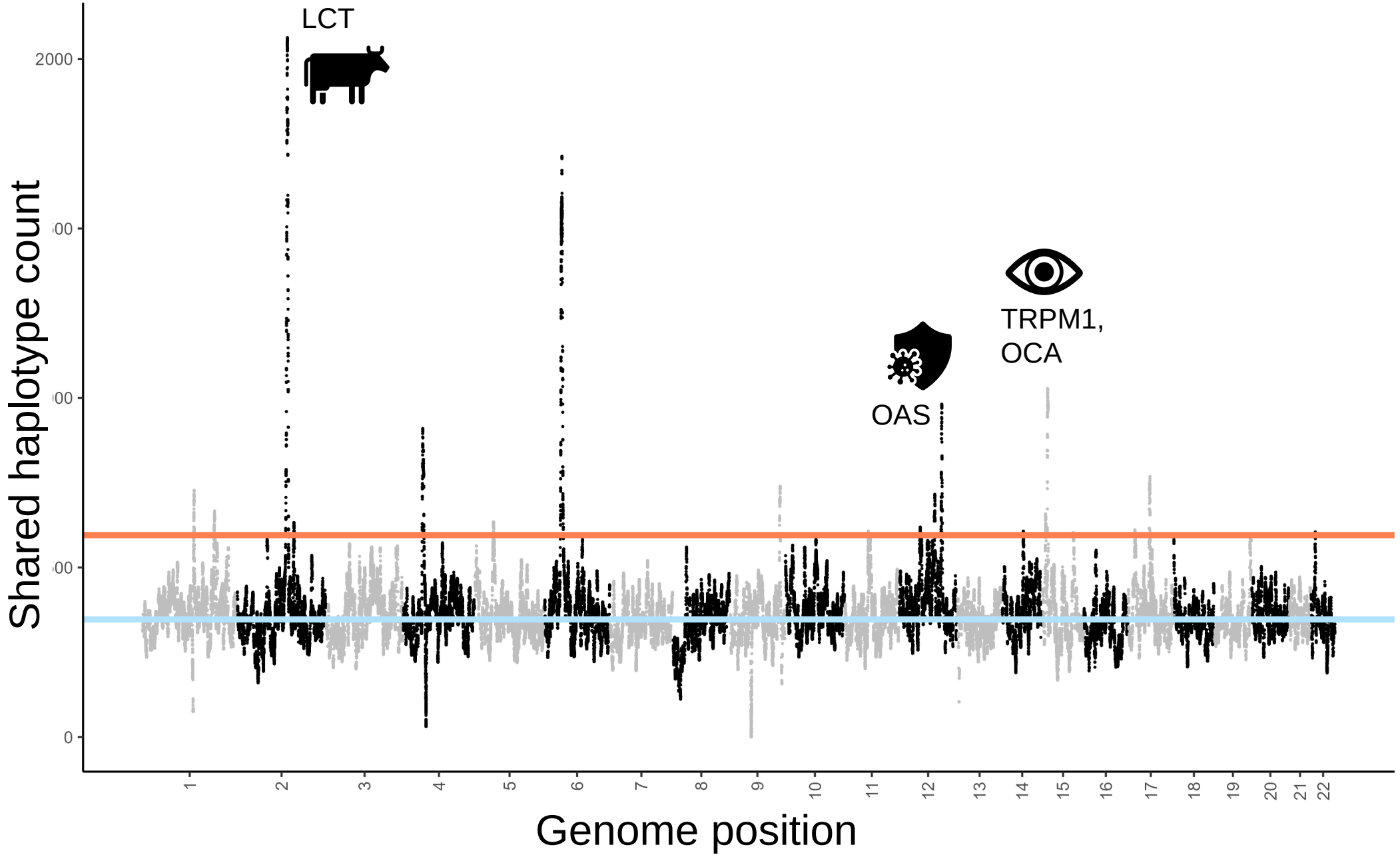
My vision in this role

- Today: teaching undergrads in public health
- Able to rigorous grad courses
- Passionate about statistical genetics
  - Broadening curriculum beyond human genomics
  - Outreach (workshops, online courses)
  - Mentoring student research
  - Promoting DEI

Thank you for having me!

Appendix: my research


Lactase persistence
Estimate
Interval
Bootstrap
Frequency
Time ago (generation)
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
250
300
0.00
0.01
0.02
0.03
0.04
0.05


LCT
TRPM1,
OCA
OAS
Shared haplotype count
2000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Genome position

Future research

- Recent selection in non-European or non-human studies
- Deep learning in population genetics
- Correlated binary random vector in light of stochastic process