HYPOTHESIS TESTING

Scott Klemmer and Michael Bernstein

Analyzing your data in 3 questions

1. **What does my data look like?**
   Explore your data graphically
   Plot all your data
   Plot several different summaries
2. **What are the overall numbers?**
   Aggregate statistics for each condition
   Usually mean and standard deviation
3. **Are the differences "real"?**
   Compute significance (p value)
   Likelihood that results are due to chance

Is my coin biased?

Null hypothesis

Scientific default skepticism: the coin is balanced
Goal: falsify the null hypothesis

How likely is 13 heads or 13 tails?

- Or even more?

# heads	Probability	# heads	Probability
0	0.00000095	10	0.17619705
1	0.00001907	11	0.16017914
2	0.00018120	12	0.12013435
3	0.00108719	13	0.07392883
4	0.00462055	14	0.03696442
5	0.01478577	15	0.01478577
6	0.03696442	16	0.00462055
7	0.07392883	17	0.00108719
8	0.12013435	18	0.00018120
9	0.16017914	19	0.00001907
		20	0.00000095

Sum the probabilities

# heads	Probability	# heads	Probability
0	0.00000095	10	0.17619705
1	0.00001907	11	0.16017914
2	0.00018120	12	0.12013435
3	0.00108719	13	0.07392883
4	0.00462055	14	0.03696442
5	0.01478577	15	0.01478577
6	0.03696442	16	0.00462055
7	0.07392883	17	0.00108719
8	0.12013435	18	0.00018120
9	0.16017914	19	0.00001907
		20	0.00000095

The sum is…

- Summed probability: p=0.263
- Thus, we'd expect 13 or more heads (or 13 or more tails) roughly 25% of the time we flip a coin twenty times
- 14 or more: p=0.11
- 15 or more: p=0.04

How low does the probability need to be for us to declare the coin biased?

Statistical significance at p=.05

one in twenty occurrences
is a scientific norm

The process in a nutshell

- Take note of our outcome, compared to a baseline
  - 13 heads out of 20 coin flips, compared to an unbiased coin
  - 200 signups out of 1000 pageviews, compared to our control interface getting 180 signups out of 1000 pageviews
  - Average of 20 photos posted per month with our new interface, compared to 19 with our old interface
- Sum the probability of all outcomes at least that unlikely
- Compare to statistical significance margin p=.05

How do we calculate the probability?

today: two statistical tests

Pearson's chi-square test

When do I use a chi-square test?

- **Chi-square compares count data**
  - “My coin produced thirteen heads out of twenty, compared to an unbiased coin that would produce ten heads.”
  - “Twenty people clicked on the banner when it was blue, vs. forty people clicked on it when it was black.”

- **Chi-square cannot compare continuous measures**
  - “The average runner with our shoes ran 18 miles.”
  - “The average time to completion with was 100 seconds with Interface A and 140 seconds with Interface B.”

Compare observed vs. expected

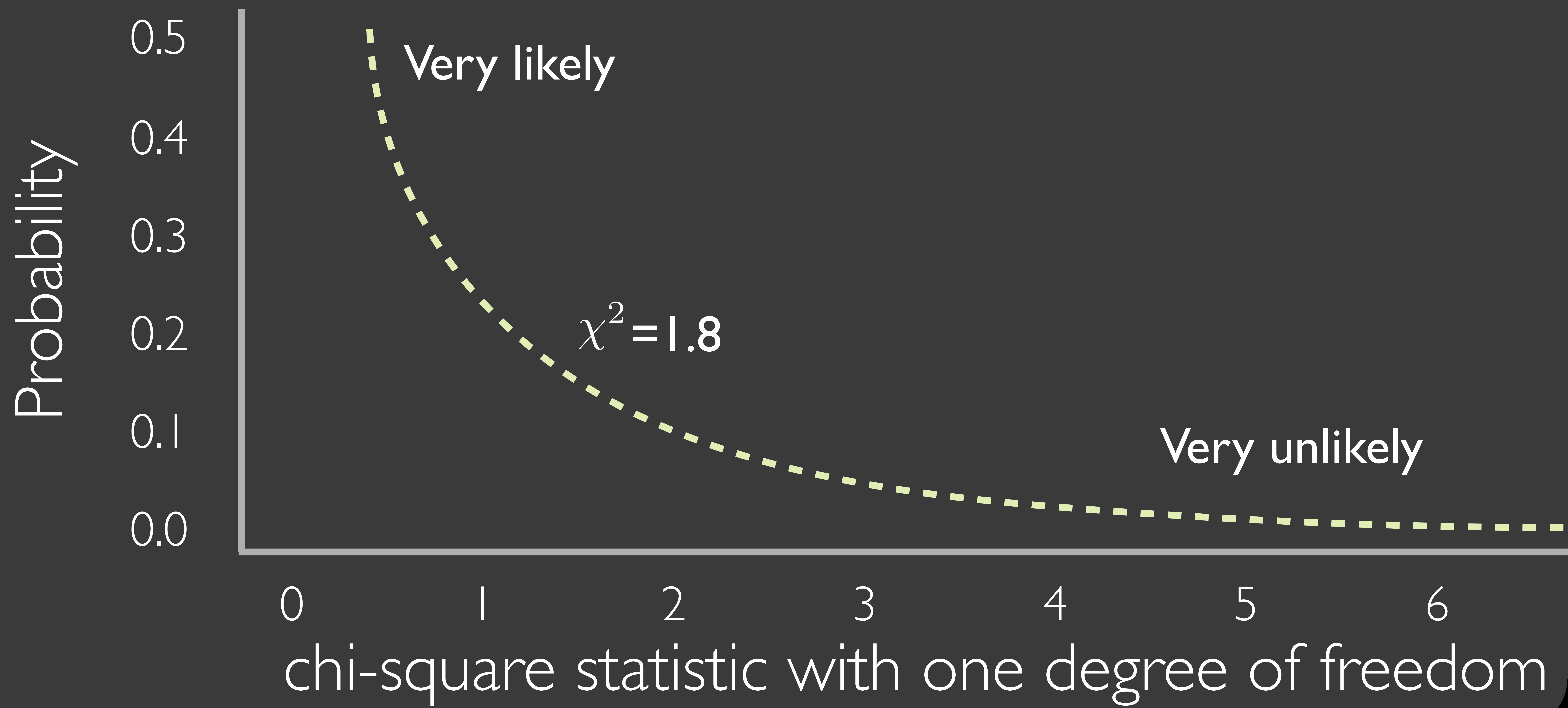
	heads	tails
observed		
expected		

Pearson's Chi-Squared statistic

$$\chi^2 = \frac{(observed - expected)^2}{expected}$$

Sum this value over all possible outcomes

These calculations produce a chi-square distribution



Calculating the chi-square statistic

- Use R or Excel

```
> pchisq(1.8, 1)
[1] 0.8202875
```

```
=CHISQ.DIST(1.8,1,TRUE)
0.82028751
```

- These calculate the value of the distribution to the left of the statistic: we need the rest.
- So, the p value is 1 - 0.82.
  p=0.18: we cannot reject the null hypothesis.

What if the trend continued?

- Say we tossed a coin 60 times, and saw the same pattern: 39 heads out of 60

	heads	tails
observed		
expected		

What if the trend continued? (2)

- What is the p-value?

```
> pchisq(5.4, 1)
[1] 0.9798632
> 1 - pchisq(5.4, 1)
[1] 0.02013675
```

- p = 0.02, so the difference is significant

Example: Improved click-throughs?

- A web site has a button labeled "sign up".
  10% of visitors click the button.
- They create an alternative, "learn more". It gets 1000 visitors and 119 conversions.
- Can we say with confidence that the "learn more" button has a higher click-through rate than the "sign up" button?

Example: Improved click-throughs?

- The odds that the observed difference happened by chance is (just barely) $p<0.05$
- The change (probably) improved click rate

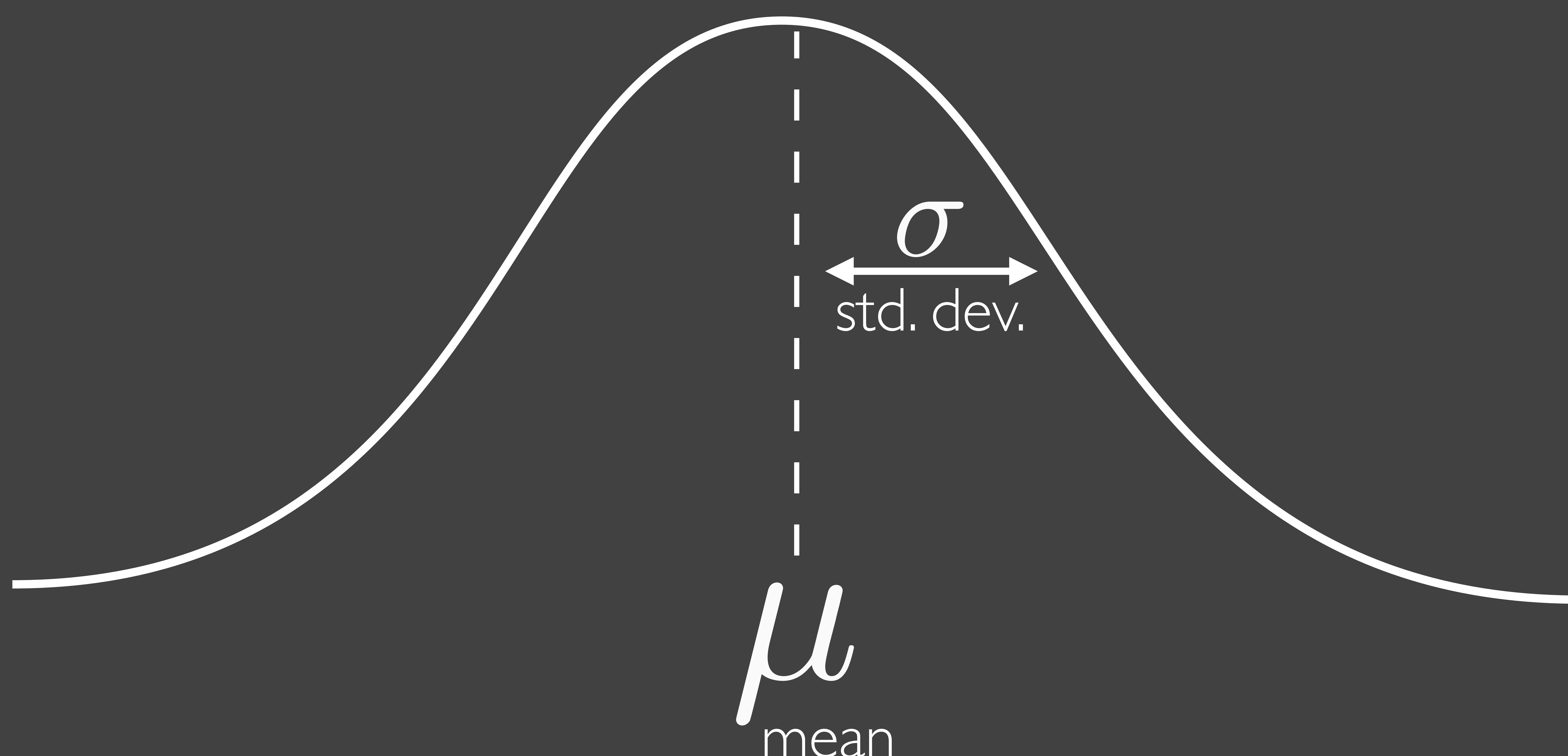
What about continuous data?

Which teaching style produces higher test scores?

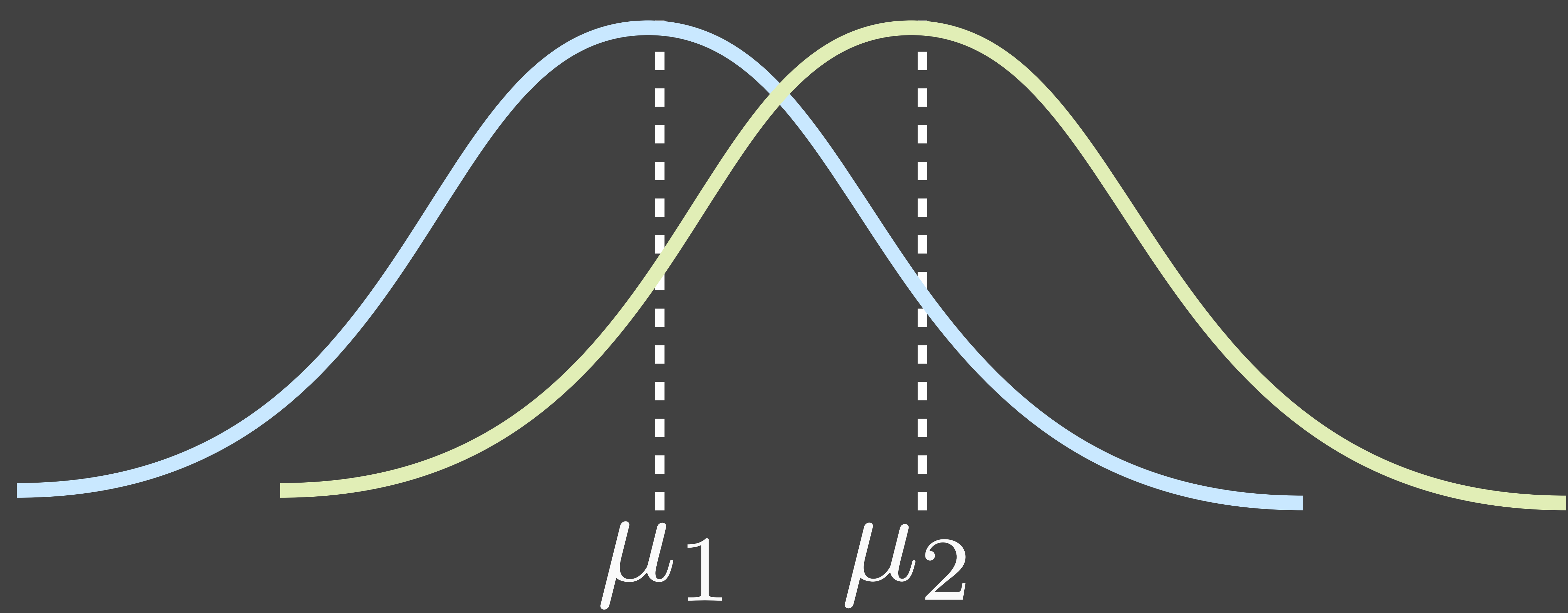
Normal Michael (control)	Hipster Michael
89pts on final exam	95
94	88
96	90
94	87
92	90
85	90
95	91
93	86
91	90
93	88

t-test

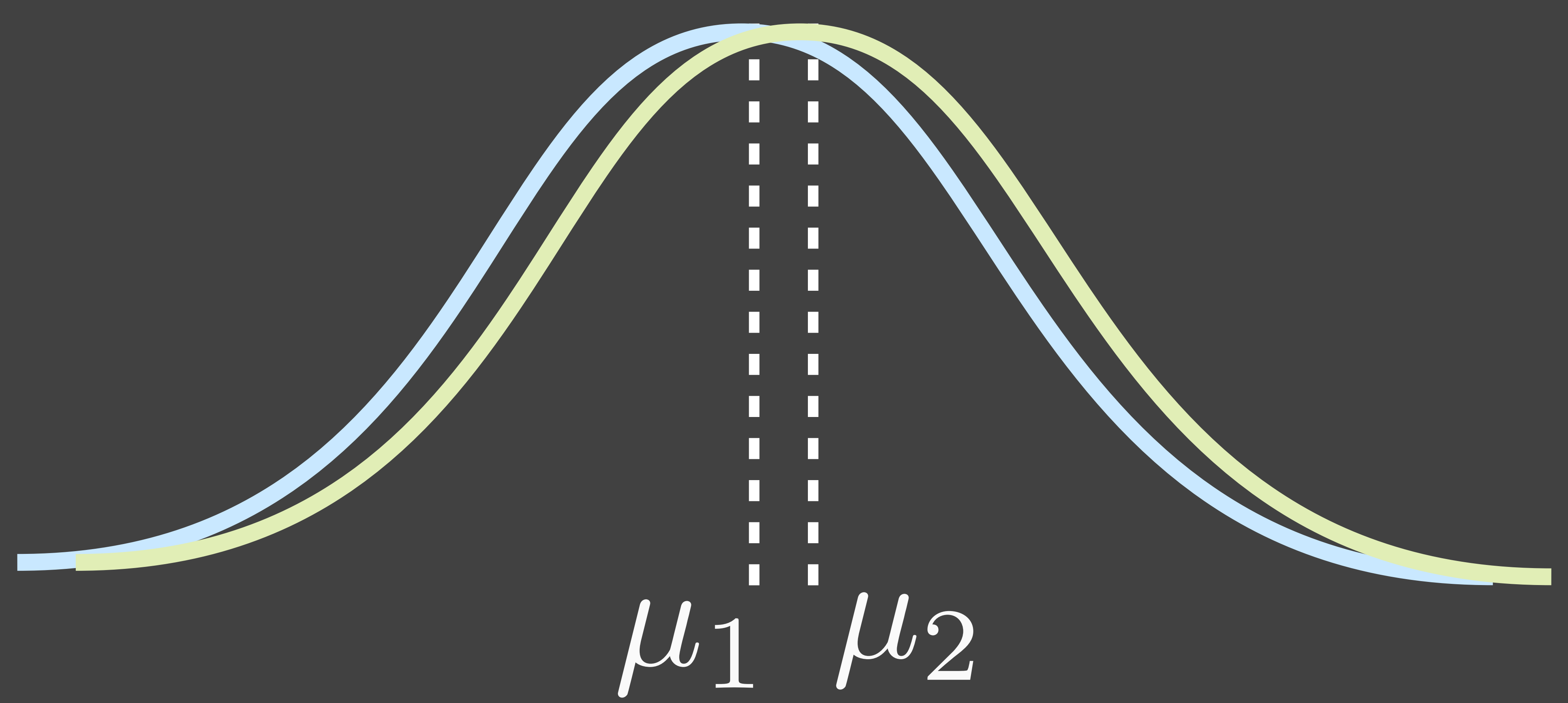
Often, continuous data is normally distributed.



t-test: do two distributions have the same mean?

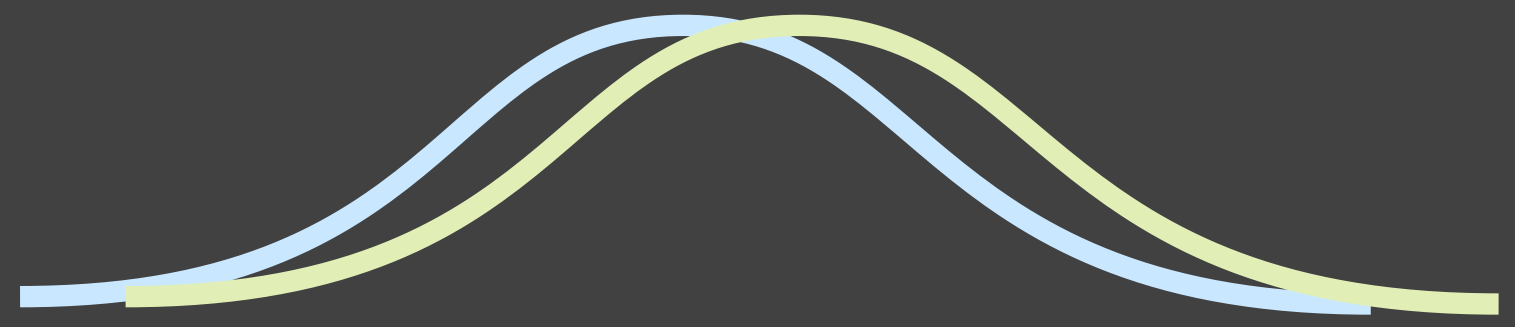


likely have different means

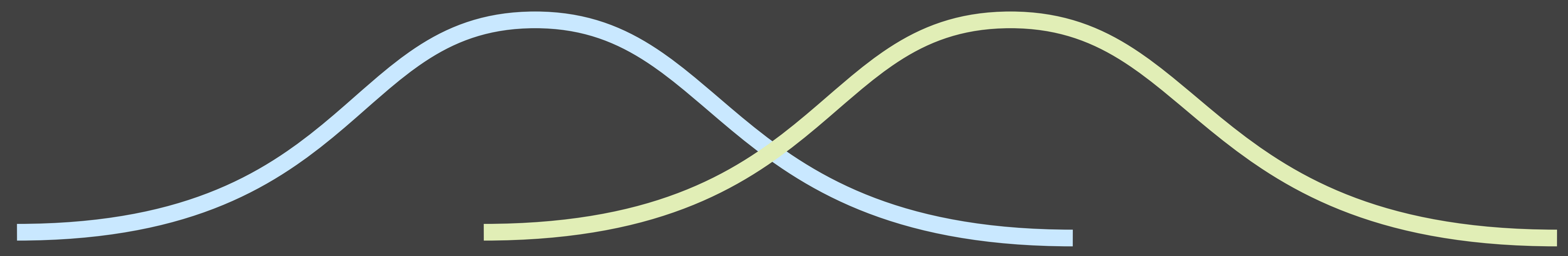


likely have the same mean
(null hypothesis)

How different are the means?



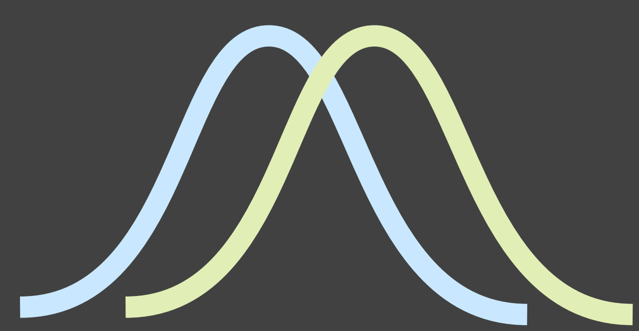
vs.



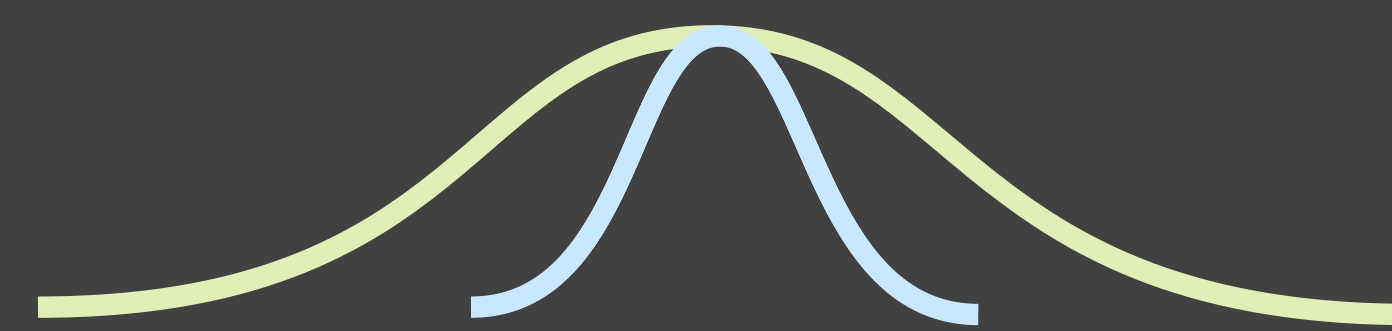
$$\mu_1 - \mu_2$$

Normal	Hipster
89	95
94	88
96	90
94	87
92	90
85	90
95	91
93	86
91	90
93	88

How similar are the variances?



vs.



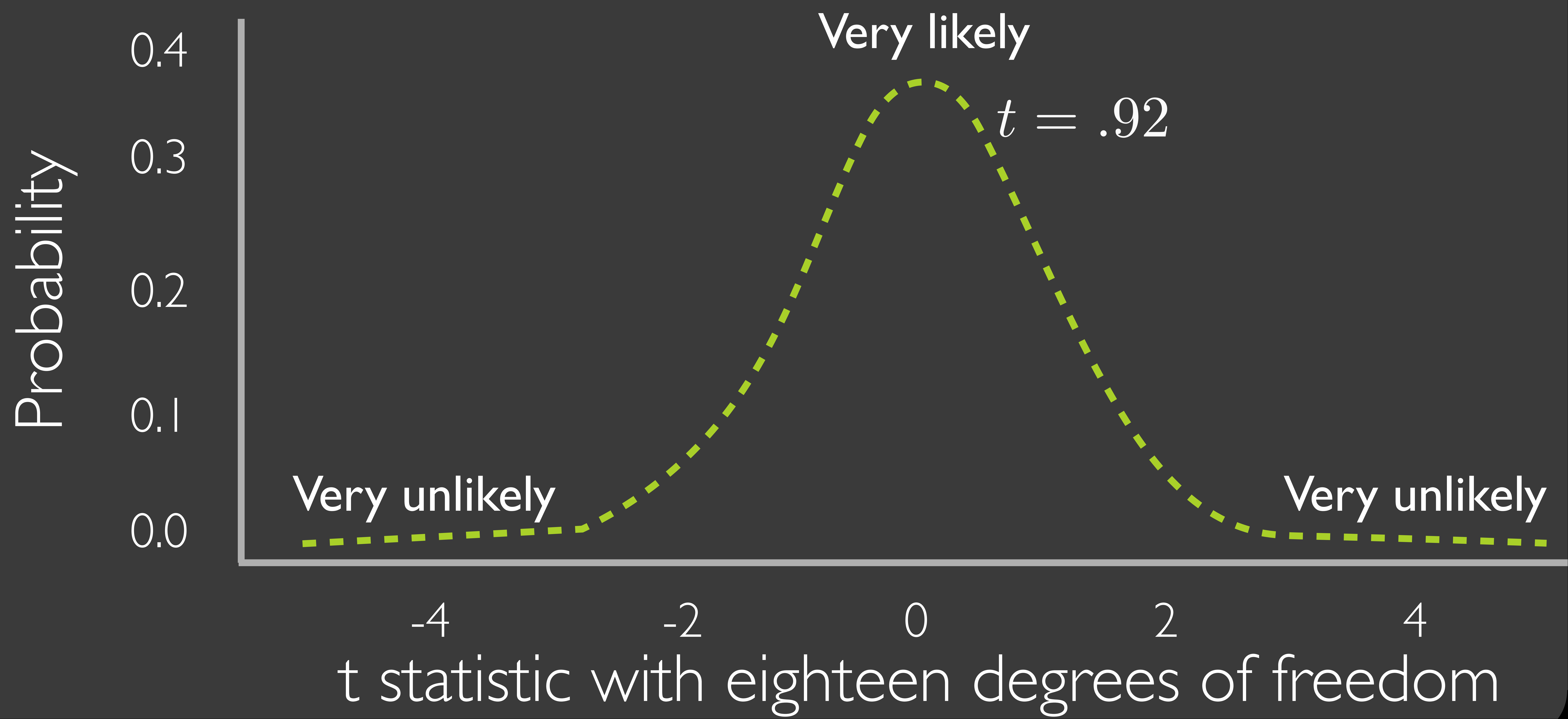
$$\frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{\square} - \frac{\sigma_2^2}{\square}}}$$

Normal	Hipster
89	95
94	88
96	90
94	87
92	90
85	90
95	91
93	86
91	90
93	88

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}}$$

Normal	Hipster
89	95
94	88
96	90
94	87
92	90
85	90
95	91
93	86
91	90
93	88
$\mu_1 = 91.5$	$\mu_2 = 90.2$
$\sigma_1^2 = 9.83$	$\sigma_2^2 = 9.96$

These calculations produce a t distribution



What are degrees of freedom?

- If we have three datapoints and we know their average, how many datapoints can vary?

$$\frac{\square + \square + \square}{3} = 5$$

Knowing the average of three numbers, we have two degrees of freedom.

So, for a t-test with two groups, we have:
$(N_1 - 1) + (N_2 - 1)$

Degrees of freedom for each test

- Chi-square: number of categories - 1
  - "If we knew the total number of observations, how many categories' counts can vary?"
  - A/B test: (2-1) = 1 degree of freedom
  - A/B/C test: (3-1) = 2 degrees of freedom

- t-test: (observations - 1) for each categories, so N - 2
  - "If we knew the average of the observations, how many observations can vary?"
  - A/B test with 100 people per condition: 98 degrees of freedom

Is the t-test significant?

- Just like the chi-square test, we need to look this up:

```
> pt(.92, 18)
[1] 0.8151308
> 1 - pt(.92, 18)
[1] 0.1848692
```

```
=T.DIST(0.92,18,TRUE)
J          K
0.81513075
```

- So p=.18, not significant

What happens if we had 4x the observations?

Before (N=20):

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}}$$

$$= \frac{91.5 - 90.2}{\sqrt{\frac{9.83}{10} + \frac{9.96}{10}}}$$

$$= .92$$

p=.18

After (N=80):

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}}$$

$$= \frac{91.5 - 90.2}{\sqrt{\frac{9.83}{\boxed{40}} + \frac{9.96}{\boxed{40}}}}$$

$$= 1.84$$

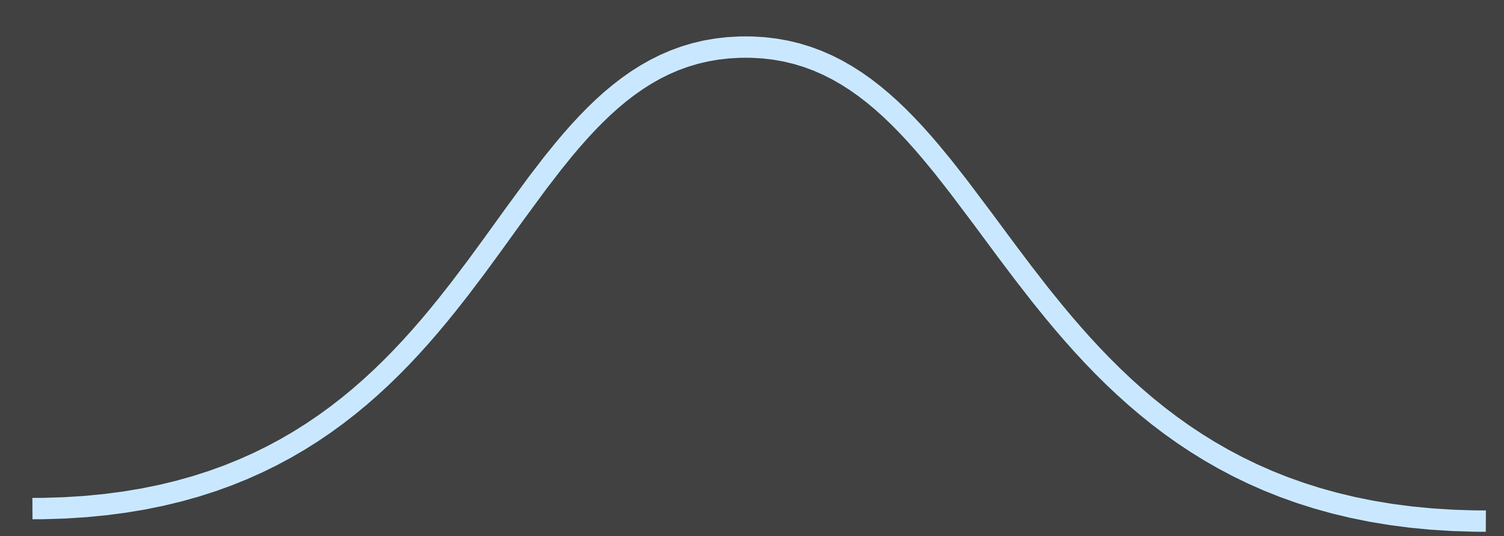
p=.03

More to learn…

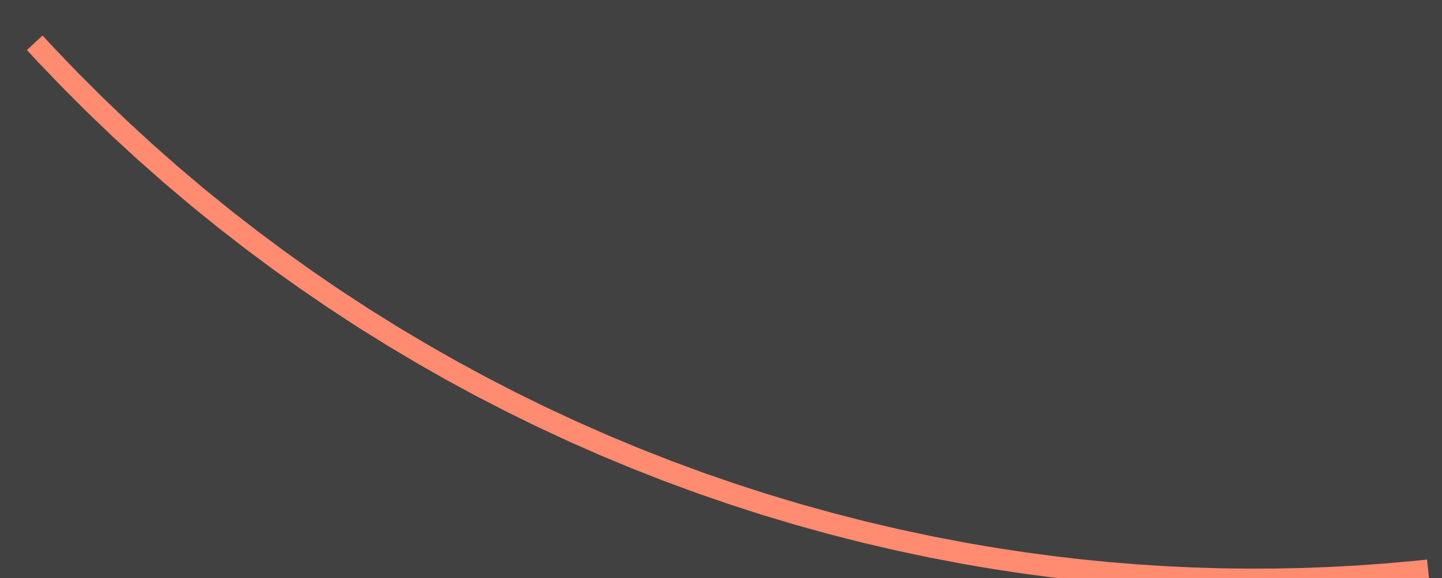
- This "unpaired" t-test is for between-subjects experiments. What if we had a within-subject experiment?

- The t-test can only handle two conditions. What if we have three or more?

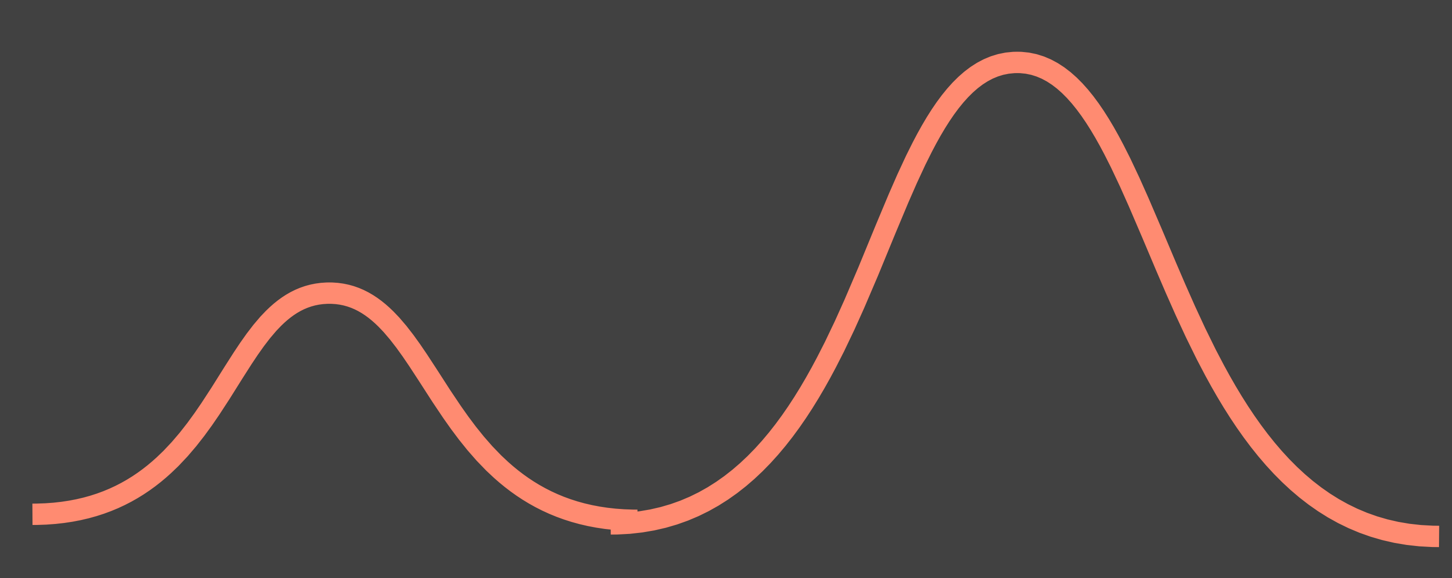
Warning: only use a t-test if the data looks roughly normally distributed



looks good



looks exponential



looks bimodal

Which to use?

chi-square test: count data
t-test: continuous data

This insight owes a lot to beer

Summary

- To get a feel for your data, graph it all
- Statistics provides tools to distinguish ‘real’ trends from ‘mirages’. It formalizes “we’re pretty sure”.
- Two common techniques:
  - For comparing rates: chi-square
  - For comparing averages: t-test