Clarifying doubts about statistics

Sakol Suethanapornkul

8 and 15 July 2020

Getting in touch...

suesakol sakolsuethana ss3302@georgetown.edu

Outline: Common ailments and how to correct them...

- **Rejection**: Statistical analysis isn't for me!
- **Putting off**: Statistics can wait until the end!
- **Misconception**: It's all about p-value, right?
  - Statistical inference
  - Sampling distribution
  - Significance testing
- **Moving forward**: It's time to think big!

About me

Research Interests:

- Statistical learning & L2/Ln acquisition
- Hierarchical probabilistic inferences
- L2 constructional knowledge & processing
- Bayesian statistics and etc.

Introduction

Symptoms of “statistics” disorder!

Do you have one or all of these symptoms?

- Become demotivated or stressed;

Symptoms of "statistics" disorder!

Do you have one or all of these symptoms?

- Become demotivated or stressed;
- Feel discomfort when thinking about t-test or ANOVA;

Symptoms of "statistics" disorder!

Do you have one or all of these symptoms?

- Become demotivated or stressed;
- Feel discomfort when thinking about t-test or ANOVA;
- Keep telling yourself, "I don't have time for this";

Symptoms of "statistics" disorder!

Do you have one or all of these symptoms?

- Become demotivated or stressed;
- Feel discomfort when thinking about t-test or ANOVA;
- Keep telling yourself, "I don't have time for this";
- Always look to hire people to run statistical tests for you;

Symptoms of "statistics" disorder!

Do you have one or all of these symptoms?

- Become demotivated or stressed;
- Feel discomfort when thinking about t-test or ANOVA;
- Keep telling yourself, "I don't have time for this";
- Always look to hire people to run statistical tests for you;
- Believe that all is needed is just a bunch of p-values.

Symptoms of "statistics" disorder!

I hope this webinar will help you feel more positive about statistics!
I'm not going to throw at you a bunch of numbers or calculations.
Instead, I'll guide you through some of the key concepts that underlie statistical analyses everyone loves (to hate)!

A broader perspective

- **Statistical expertise** and **methodological rigor**;
  - Improve the quality of research in Thailand
- **Quality research** and **societal impact**;
  - Ask what your M.Ed. degree can do for your country
- Tools for **professional development**
  - Be informed and help other people become more informed

Downright Rejection

Point 0: Statistical analysis isn't for me!

As part of data analysis, statistics is a **tool** that can give you valuable information about patterns of data, whether you are a qualitative or quantitative person. :-)

Point 0: Statistical analysis isn't for me!

Types of questions in data analysis

- **Descriptive**: summarizing characteristics of data at hand
- **Exploratory**: exploring unknown relationships between variables
- **Inferential**: using data collected to make statement about population
- **Predictive**: using current or existing data to predict future data
- ....and a few more.

Source: Six types of data questions

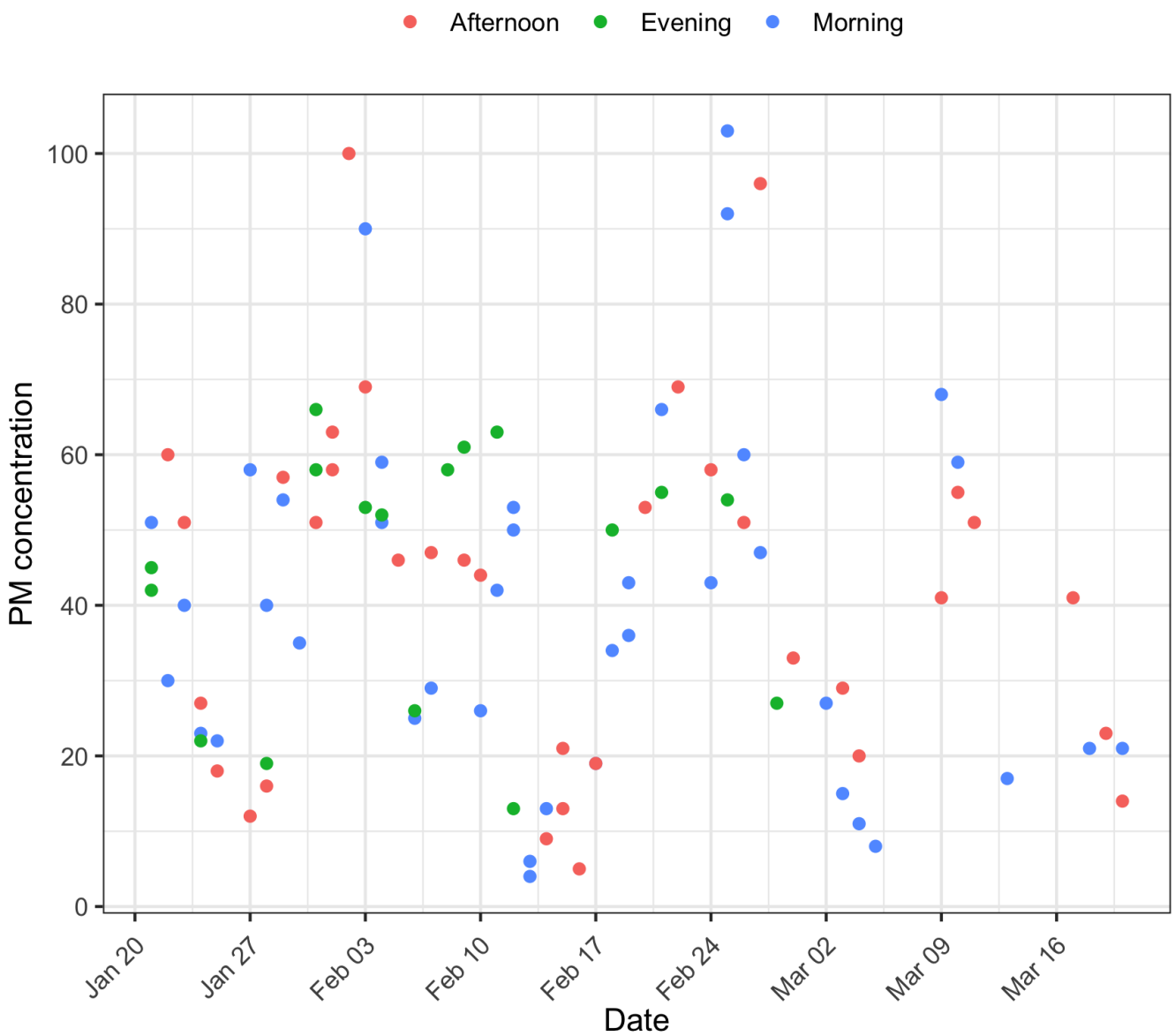
Exercise 0.1

Exploratory question: PM 2.5 concentration in Bangkok from January to March 2020

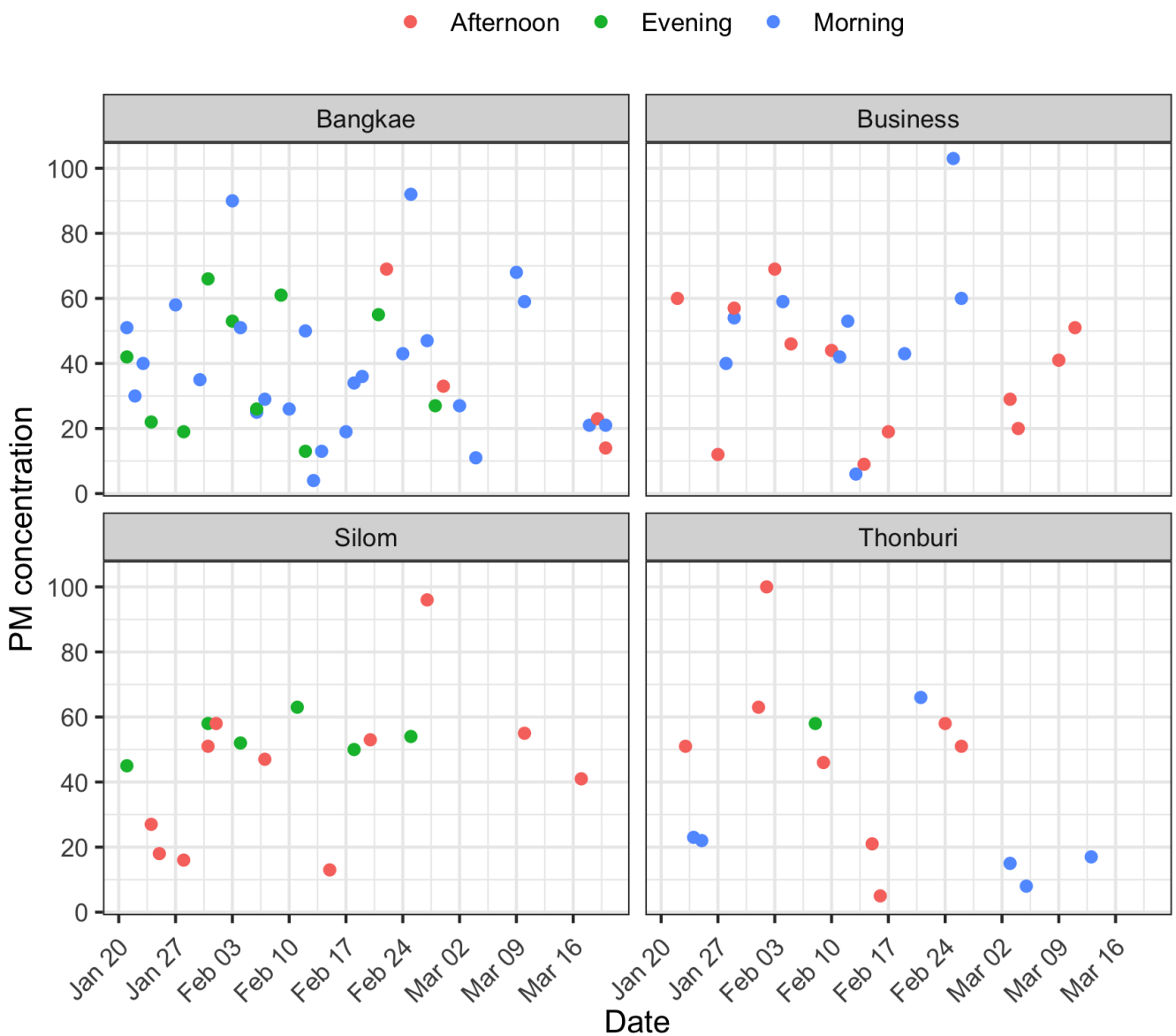
- PM 2.5 = tiny particles that float in the air ($\leq 2\frac{1}{2}$ microns)
- concentrations recorded from a small, portable air detector

Instructions: consider whether there's any relationship between PM concentrations and select variables. If you think there is, what kind of relationship do you see?

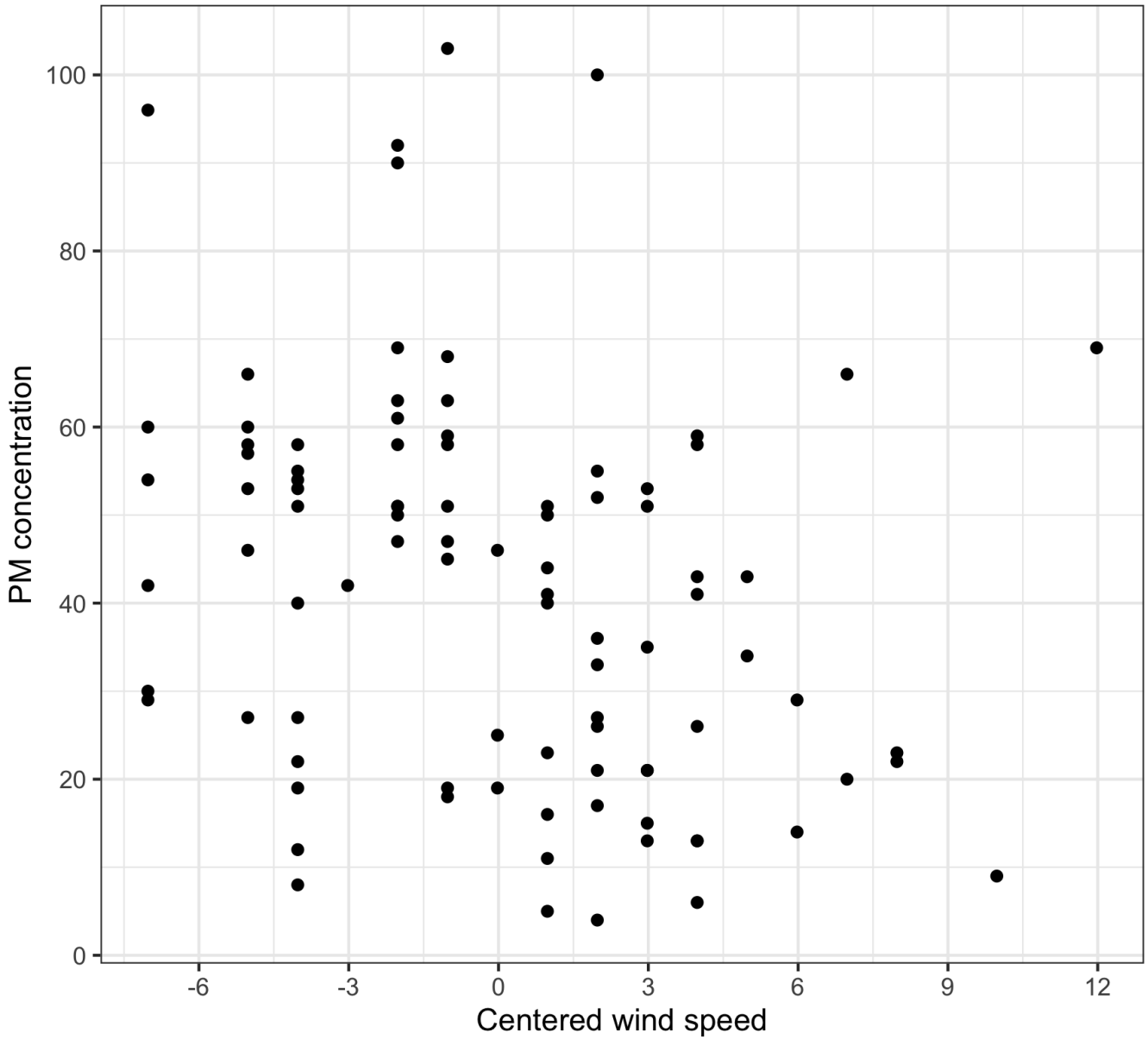
Exercise 0.1


Afternoon
Evening
Morning
PM concentration
100
80
60
40
20
0
Jan 20
Jan 27
Feb 03
Feb 10
Feb 17
Feb 24
Mar 02
Mar 09
Mar 16
Date

Exercise 0.1


Afternoon
Evening
Morning
Bangkae
Business
Silom
Thonburi
PM concentration
0
20
40
60
80
100
Jan 20
Jan 27
Feb 03
Feb 10
Feb 17
Feb 24
Mar 02
Mar 09
Mar 16
Date

Exercise 0.1


PM concentration
100
80
60
40
20
0
-6
-3
0
3
6
9
12
Centered wind speed

Discussion 0.1

1. Do PM concentrations pattern with any of the variables? What are the variables?
2. In the last figure, what kind of relationship between wind speed and PM concentrations did you see?

Point 0: Bite-size summary

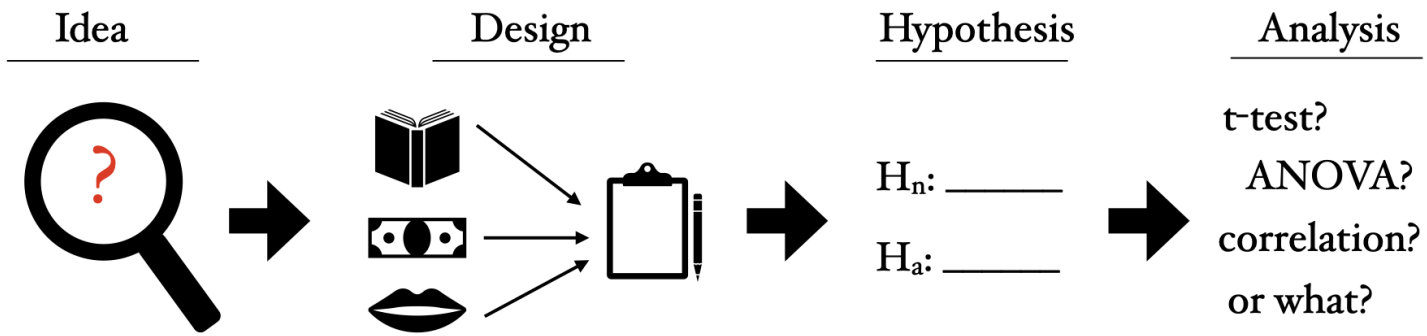
Even though you are not planning to conduct a **quantitative** study:

1. Being able to summarize and graph data (i.e., descriptive & exploratory) is crucial; and
2. Having skills to interpret graphs and comprehend some inferential tests can be extremely useful. :-)

Postponing Statistical Analysis

Point 1: Statistical analysis can wait!

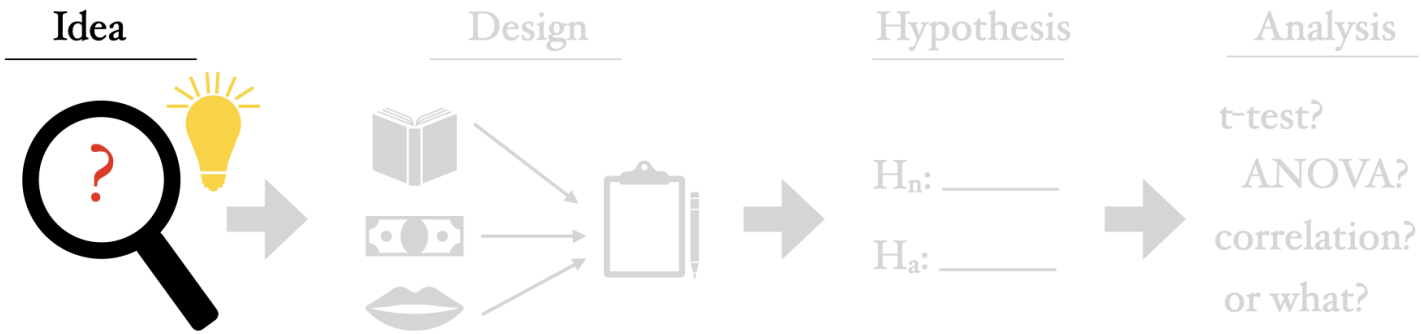
A usual research workflow(?):



In reality, things get messy, and your workflow may completely be flipped! Don't despair! You're not alone.

Point 1: Statistical analysis can wait!

How can I make this workflow a reality?



As you're developing and considering your research ideas, use the following points as your guide.

Point 1: Statistical analysis can wait!

Outcome variable

- Vocabulary scores
- Pass/Fail
- True/False
- Self-perceived rating

Issues

- Type (dichotomous or continuous)
- Obs (1 or > 1 per subj)
- Cluster
- Data-generating process, e.g., $y_i \sim N(\mu_i, \sigma_e)$

Point 1: Statistical analysis can wait!

Explanatory variables

Issues

- Number (i.e., how many?)
- Type
- Relationship to outcome*
- level (e.g., student- or class-level)

* In a majority of cases, we assume a linear relationship where an outcome variable, y_i, is predicted from a combination of explanatory variables.

Exercise 1.1

Instructions: In each of the following cases, answer the following questions:

1. What is an outcome variable?
   - What type?
   - Is there a cluster? How many observations per cluster?
2. What are explanatory variables?
   - What type?
   - At what level is each of these variables?

Bonus What analysis is most appropriate?

Exercise 1.1

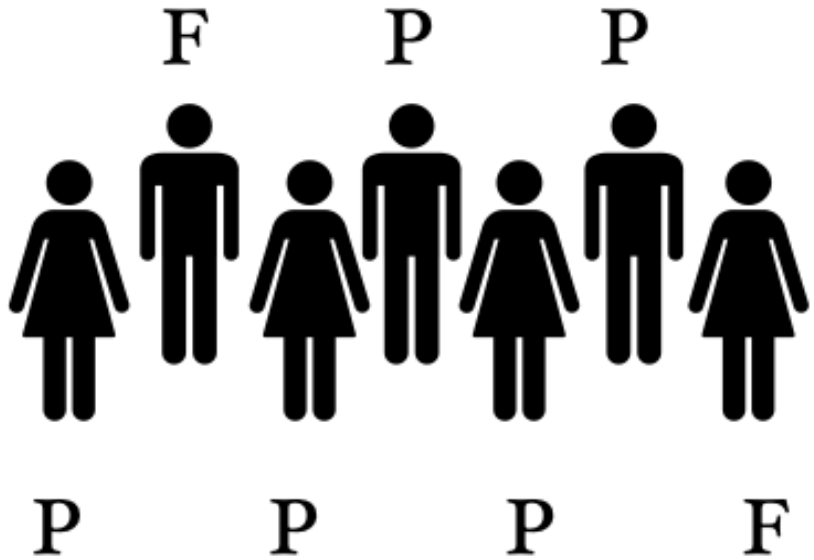
[1] We have standardized English test scores (possible range: 0–100) from 10000 students in 100 schools. In each school, 100 students are selected. We would like to investigate factors that influence test scores. We collect as part of our research design the following variables: (1) students' gender (female, male, non-binary/third-gender), (2) family SES, (3) out-of-class English use, (4) school achievement, and (5) school size.

Exercise 1.1

[2] We conduct an intervention program for 200 randomly-selected, at-risk students in one high school. After the program is terminated, we record if each student pass or fail an English course. Assuming that students are equally motivated and have roughly the same level of English proficiency, we record the following information: (1) number of hours students spend reviewing class materials, (2) students' perceived engagement, (3) gender, and (4) number of English books they read per week. We would like to know if each of these factors predict students' achievement.

Point 1: Statistical analysis can wait!

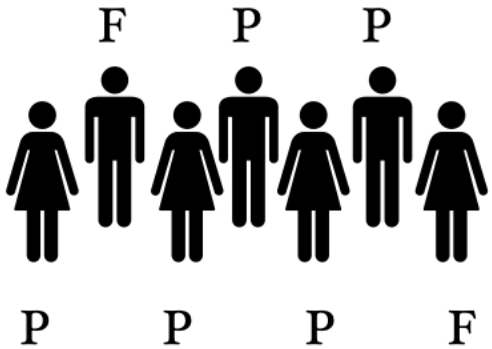
Observations from sample:



Student is at the lowest level of this model.

Point 1: Statistical analysis can wait!

Information pertaining to explanatory variables:



S1: non-binary; 2.5 hrs.; 2 books; 5/10

S2: male; 4 hrs; 4 books; 8/10

S3: female, 3 hrs; 3 books; 7/10

.....

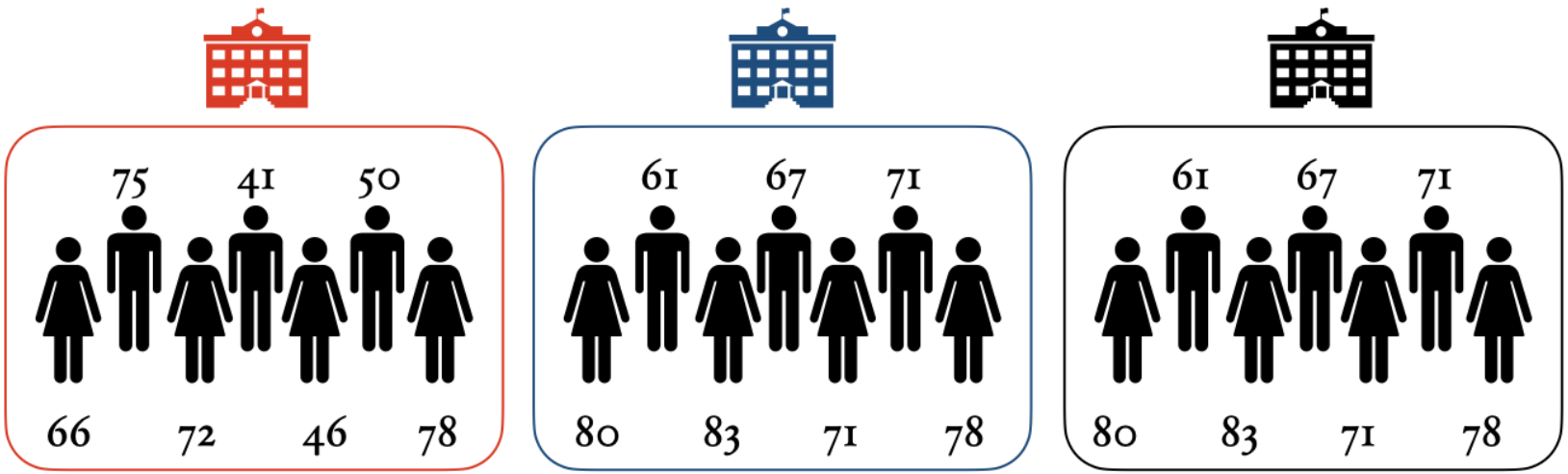
Point 1: Statistical analysis can wait!

Bonus: Setting up a spreadsheet in Excel

ID	Gender	Eng_hrs	Books	Enjoy	Success
1	Non-binary	2.5	2	5	Pass
2	Male	3	3	7	Fail
3	Female	4	2	6	Pass
4	Non-binary	5	4	8	Fail
5	Female	4	1	9	Pass

Point 1: Statistical analysis can wait!

Contrast the previous example with this one, where students are "nested" in schools.



Point 1: Statistical analysis can wait!

You also have explanatory variables related to students and schools.

School 1: 4000 students; A grade

S1: male; 300k; 3 hrs.

S2: female; 450k; 4 hrs.

S3: non-binary; 400k; 3..

.....

.....

Point 1: Statistical analysis can wait!

Bonus: Setting up a spreadsheet in Excel

ID	Gender	Fam_SES	Eng_use	School_size	School_achieve	Test
1	Male	300	3	4000	A	75
2	Female	450	4	4000	A	68
3	Non-binary	275	3	4000	A	71
...	...	...	...	...	...	...
10	Female	350	5	2000	B	59
11	Non-binary	300	2	2000	B	62
12	Male	500	2	2000	B	68
...	...	...	...	...	...	...

Point 1: Statistical analysis can wait!

This isn't just about students nested within schools...

- You test vocabulary knowledge of 60 students and obtain scores on 5 areas of vocabulary. So, scores are nested within participants.

Point 1: Statistical analysis can wait!

This isn't just about students nested within schools...

- You test vocabulary knowledge of 60 students and obtain scores on 5 areas of vocabulary. So, scores are nested within participants.
- You focus on English test achievement of 200 schools in 20 school districts. So, schools are nested within districts.

Point 1: Bite-size summary

Before you go out and collect data:

1. Know your outcome and explanatory variables; and
2. Have a clear idea as to how the two kinds of variables are related.

Some misconceptions

Point 2: There's one way to conduct statistical inference

A common refrain you may hear (or use yourself):

- Focus on a p-value of the test you run and make sure it is significant (i.e., $p < 0.05$);
- The smaller the p-value, the stronger the evidence for your argument;
- Just run a t-test or whatever and report it! Every test is the same!

Upshot: These ideas are all bad and dangerously wrong!

Point 2: There's one way to conduct statistical inference

To keep your anxiety level low, we will focus narrowly on:

- approaches to statistical inference (frequentist and Bayesian);
- inferences one can conduct in frequentist settings (i.e., point estimate, significance testing, etc.)

But first, a quick word about **statistical inference**

Point 2: There's one way to conduct statistical inference

Statistical inference (with inferential statistics): a process by which statements about *an underlying probability distribution of data* are made.

More on this in a minute...

Point 2: There's one way to conduct statistical inference

Random experiments and random variables:

When you toss a coin or measure a person's blood pressure, you do not know what an outcome will be. This kind of process is referred to as a **random experiment** in statistics.

A random experiment generates random outcomes. And a **random variable** is a **numerical description** of each outcome of the random experiment. Random variables are written with capital letters such as X or Y

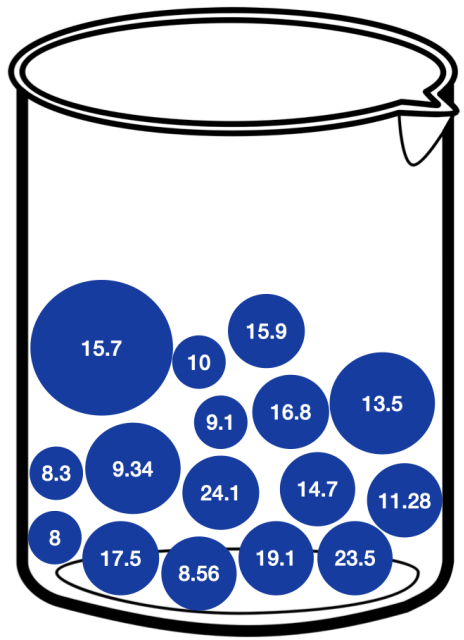
Point 2: There's one way to conduct statistical inference

Examples	
Coin toss	H $\rightarrow$ 1; T $\rightarrow$ 0
Sex at birth	M $\rightarrow$ 1; F $\rightarrow$ 0
Enjoyment	Unhappy $\rightarrow$ 1; OK $\rightarrow$ 2; ...
Height, RTs,...	$\mathbb{R} := [0, ...]$

Every (discrete or continuous) random variable has associated with it a probability distribution. This is also true of your outcome variable (e.g., vocabulary scores, motivation profiles, etc.) It can be associated with a probability distribution!

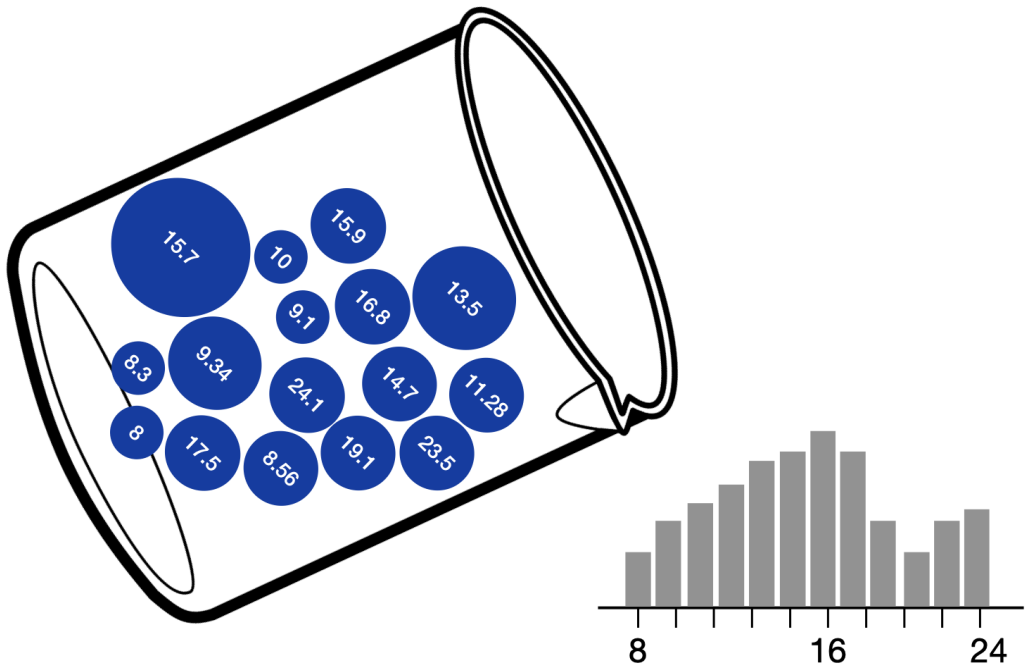
Point 2: Probability

Let's say we have a measure of speech error ($[\frac{\text{errors}}{\text{total words}}] \times 100$).



Point 2: Probability

Let's say we have a measure of speech error ($[\frac{\text{errors}}{\text{total words}}] \times 100$).



Example 2.1

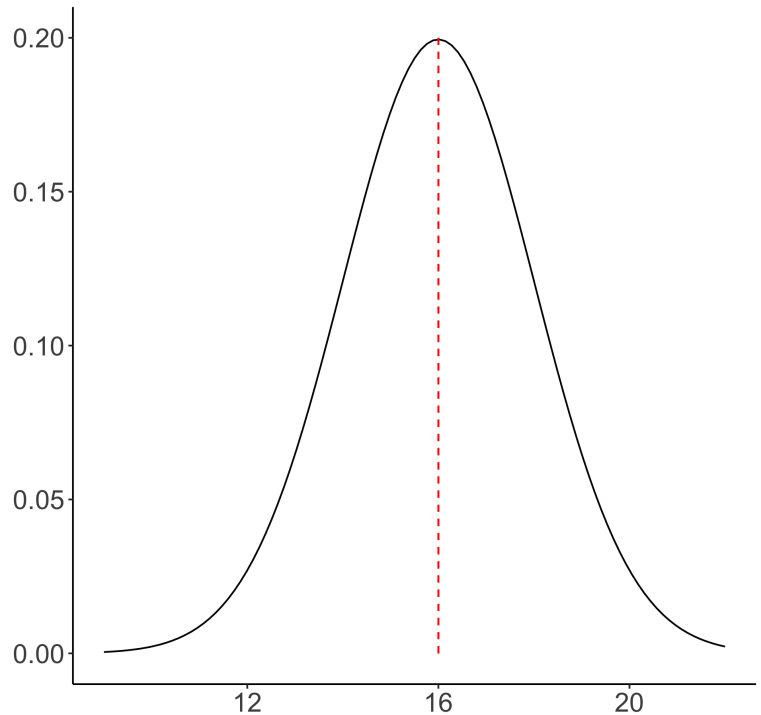
Because our "speech error" is a random variable and because it has a probability density function*, we can make a theoretical probability distribution of the outcome variable and obtain a lot of information from it!

Let's just say our $\bar{X} = 16, SD = 2$.

Just so you know, there are important characteristics of random variables that unfortunately we don't have time to discuss.

Example 2.1

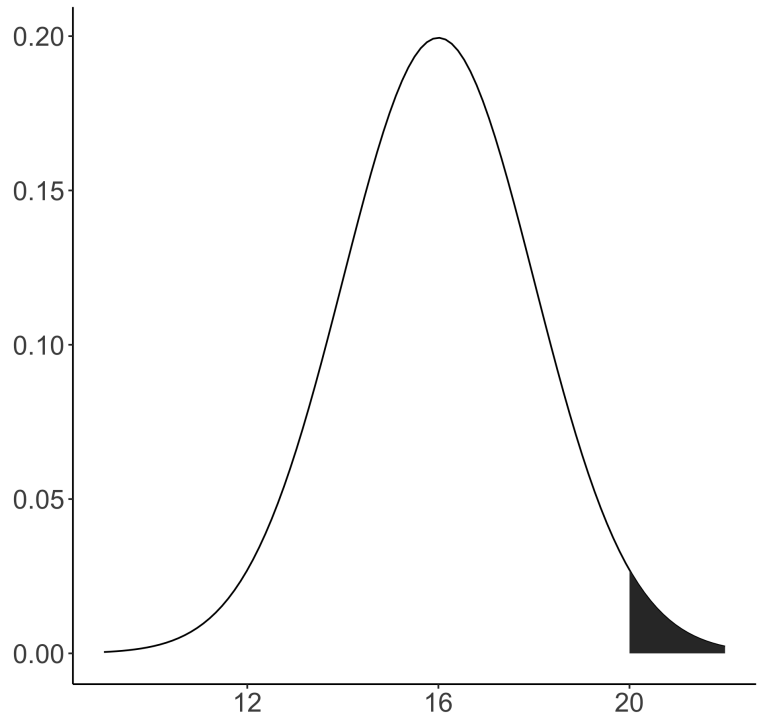
Assuming that these values are also our population parameters*, $\mu = 16, \sigma = 2$, we get:



More on this "population" thing later... :-)

Example 2.1

You might be wondering why we're even discussing this...
We can find out the probability of a value (say ≥ 20) under this curve. $p(X \geq 20) = 0.023$.



Example 2.1

Looking ahead: In significance testing ($p < 0.05$ or $p > 0.05$), a **theoretical distribution of Null Hypothesis** (H_0) is constructed and test statistic is checked against that distribution!

If the probability of H_0 "producing" that value is lower than, say, 0.05, we can then reject H_0.

We will see more examples of this along with some visualization. Our goal is to correct mistakes that most people make regarding the p-value.

Point 2: Probability

Like everything else, there is more than one interpretation of probability.

	Relative frequency	Subjective
Idea	Probability of some event A, p(A), is obtained from infinite repetitions; Probability is frequency of occurrence	Probability is a degree of belief a person has about an event;
Method	Frequentist statistics	Bayesian statistics

Point 2: Statistical inference

What I said: Statistical inference is a process by which statements about *an underlying probability distribution of data* are made.

What I should have said: As we conduct statistical inference, our goal is to use data from a "random" sample of size *n* to learn about the distribution of unknown parameters , denoted by Greek letters (e.g., μ, θ, β etc.)

Example 2.3: Effect of perceptual training

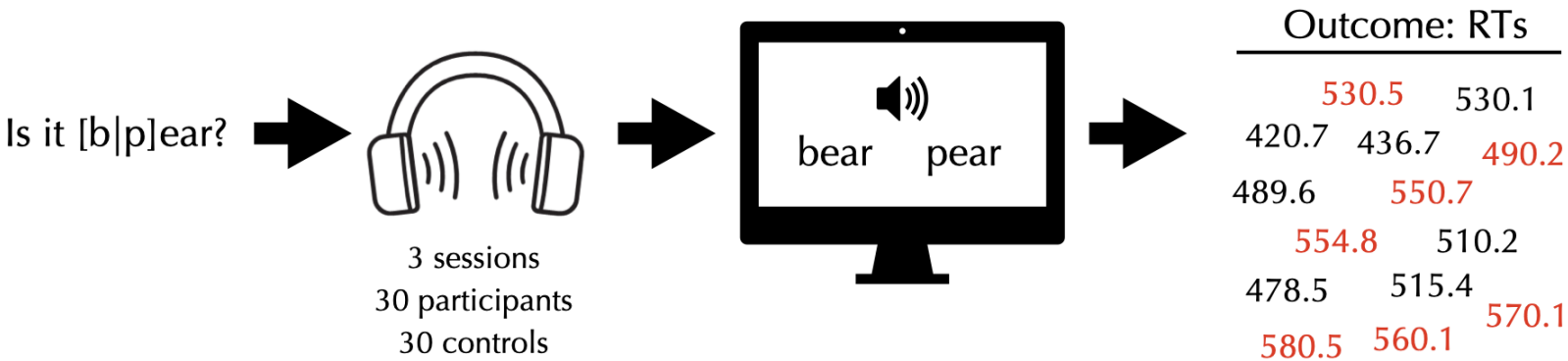
Is it [b|p]ear?

Example 2.3: Effect of perceptual training

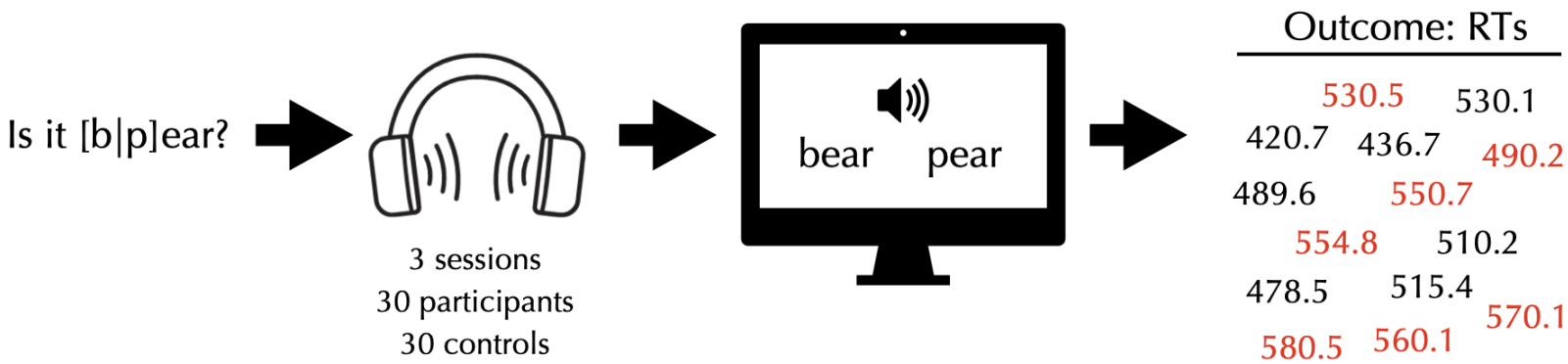
Is it [b|p]ear?

3 sessions
30 participants
30 controls

Example 2.3: Effect of perceptual training



Example 2.3: Effect of perceptual training



Data & sample statistics:

$\text{RT}_{\text{Exp}} = \{530.1, 420.7, 436.7, 489.6, 510.2, 478.5, 515.4\}$ and

$\text{RT}_{\text{Con}} = \{530.5, 490.2, 554.8, 570.1, 580.5, 560.1\}$

Example 2.3: Effect of perceptual training

In this example, you want to learn about "the effect of training on RTs,"

- You posit there is a true effect of such perceptual training in the population (say, θ);
- As a parameter with some distribution, θ gives rise to the data we have observed;
- But because θ is unknown, we must use the data from our sample to estimate it.

This is **point estimation**, which is separate from **significance testing** (i.e., obtaining a *p*-value of a test statistic). More on these two later...

Point 2: Statistical inference

Frequentist

- Parameters are unknown, fixed constants;
- Probability statements can't be assigned to parameters;
- Inference is conducted with a **sampling distribution**.

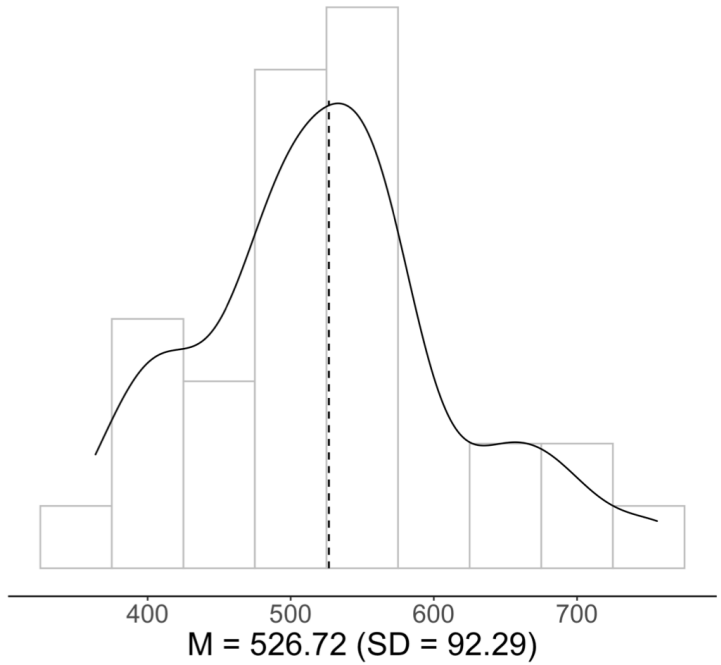
Bayesian

- Parameters are unknown but assumed to be random variables;
- Probability statement is given to parameters;
- Inference is conducted using Bayes' theorem.

Point 2: Sampling distribution

We can calculate sample statistics from a **single sample**, namely $\bar{x}$ and SD. For now, we will focus on the experimental group (n = 30). Suppose we have the following data:

542.72	481.85	646.92	556.48	573.04
755.87	541.80	479.10	469.72	506.64
507.21	418.19	568.53	675.74	401.84
431.98	509.03	571.62	560.77	694.13
524.42	392.63	404.82	556.22	474.43
548.28	363.61	634.87	494.34	514.91

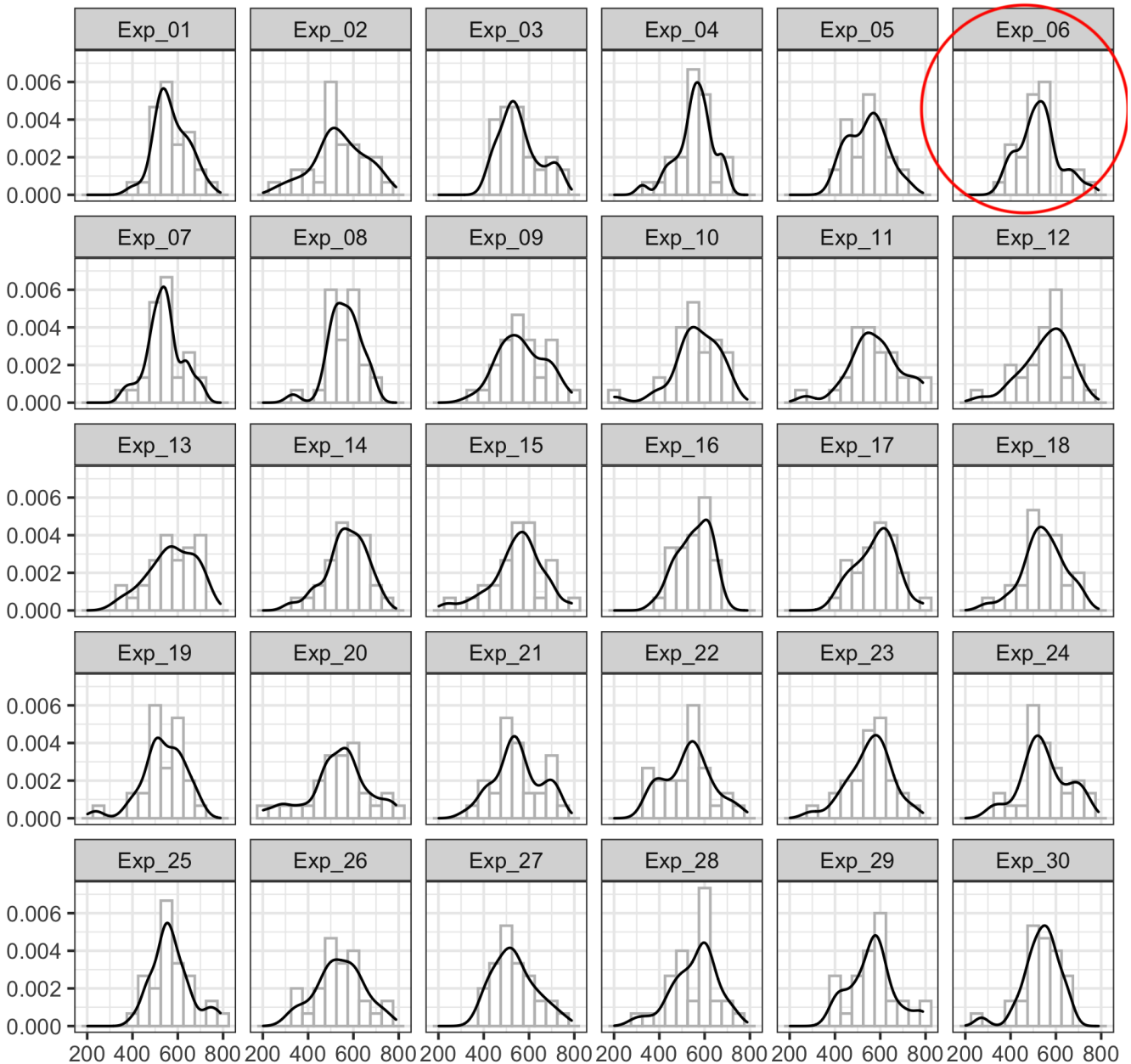


Point 2: Sampling distribution

Imagine that the true RTs of our experimental group is 550 ms. Note that this is *not* the true effect of perceptual training (i.e., how much faster does the experimental group becomes).

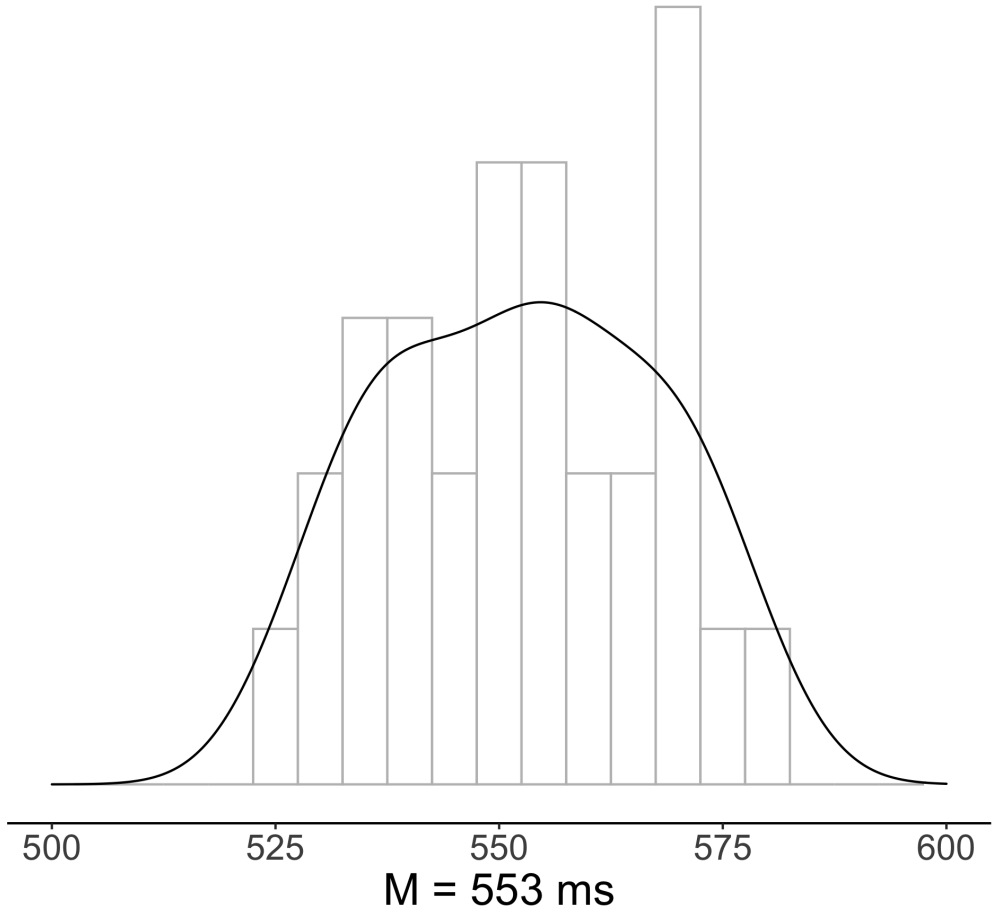
- The mean of our sample is 526.72 ms, which is not quite the true RTs;
- Suppose we have time and money. We choose to repeat this exact same experiment 29 more times (total n = 30).

Point 2: Sampling distribution



Point 2: Sampling distribution

We find $\bar{x}$ of each sample, plot those values, and calculate the grand mean of the 30 samples. Here is what we get:



Point 2: Sampling distribution

In this example, we obtain a sampling distribution of the mean. We can construct a theoretical sampling distribution for any statistic.

- Through the process, we are able to get the grand mean that is extremely close to the true population parameter value (i.e., $\mu = 550$ ms);
- If we were to repeat this process a few more times, we would land on 550 ms!

Point 2: Bite-size summary

This is a lot! If all of these things confuse you, take time to digest the materials. Statistics is a life-long pursuit. :-)