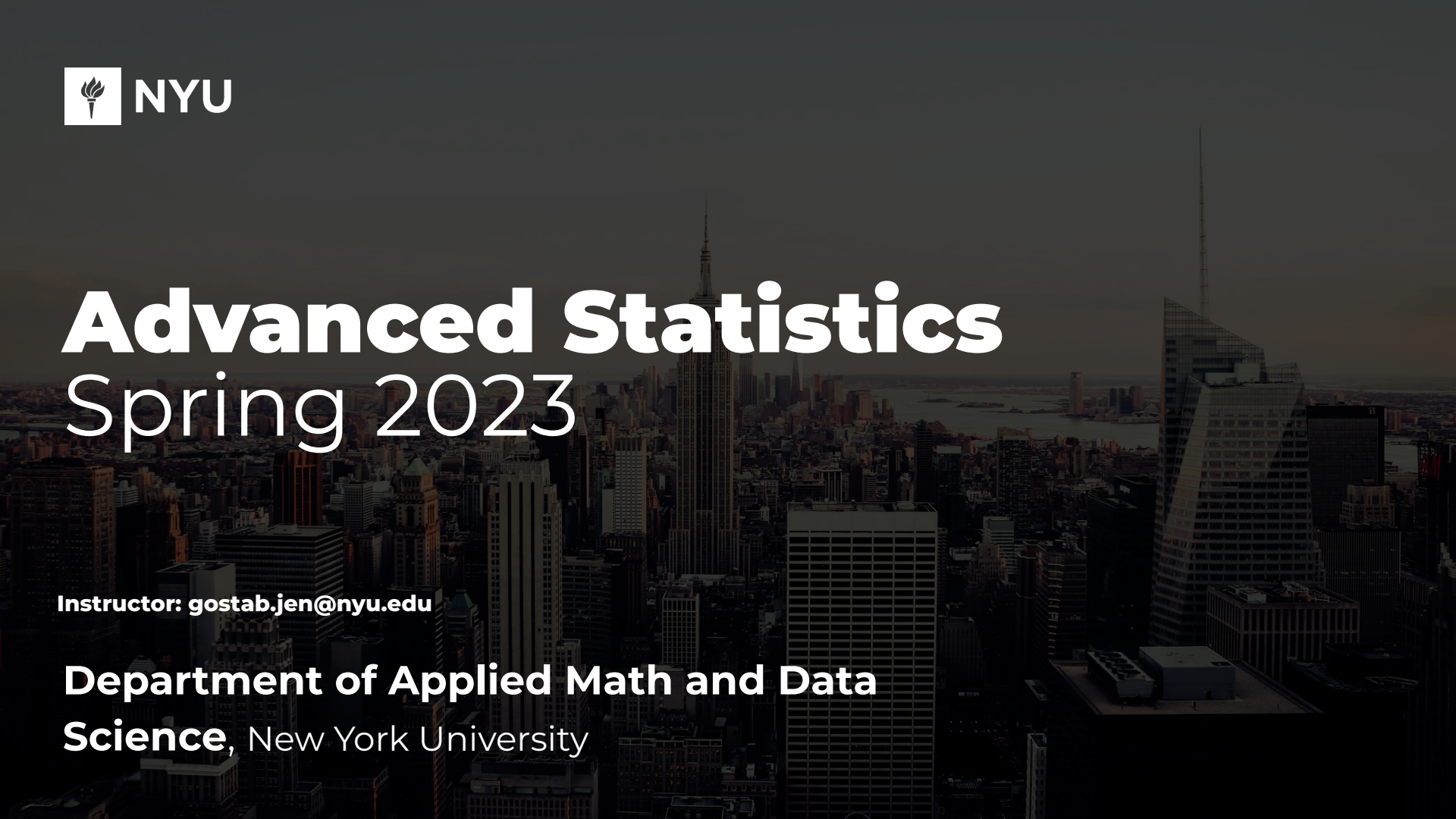


NYU

Advanced Statistics
Spring 2023

Instructor: gostab.jen@nyu.edu

Department of Applied Math and Data Science, New York University

Homework 1

- Avoid typos
- Write clear
- Label the question
- Week -4

NYU

HW Rule

See Syllabus about late submission and other requirements

Don’t forget to submit your lab activity to brightspace and ensure your full credit.

NYU

Lab Assignment

Don’t forget to submit your lab activity to brightspace and ensure your full credit

NYU

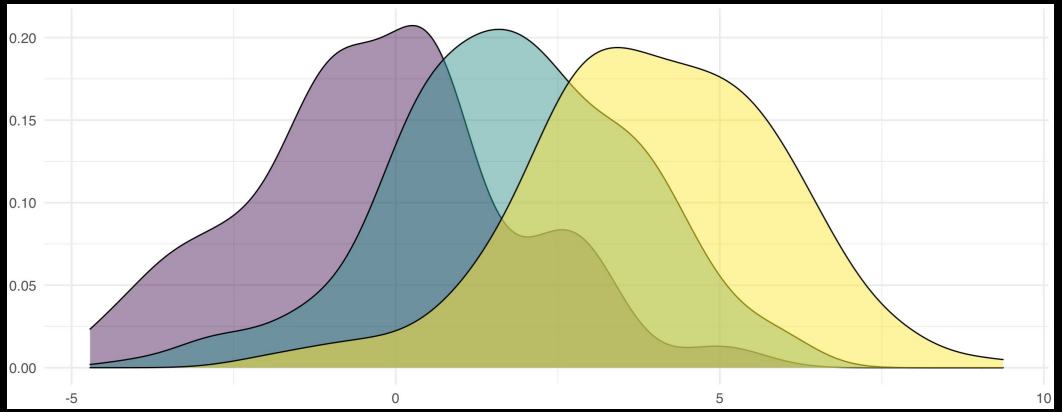
Exam

- Your exams will not involve the kind of lengthy hand calculations that you did in PSYCH-10
- Yay!
- But that requires that you have an understanding of the underlying theory behind statistics
-
-

Week 4 Plans

Review: Practice for sample exam

→Practice for R:
estimator : variance, correlation



NYU

Estimator-R exploration

check your current environment: `getwd()`

Install some key packages: `install.packages('')`

Call the library : `library()`

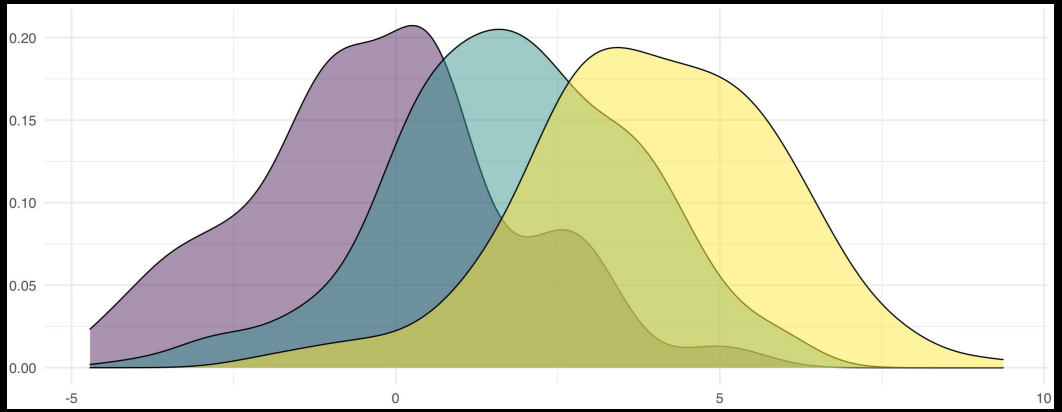
You need `arm, dplyr,ggplot2` **minimum**

Lab Activity: submit to Brightspace your function for correlation and design a test if you get the same result as cor()

```
dat <- read.csv('')
```

Descriptive Statistics

→Review: Practice for sample exam



NYU

Sample Questions for Midterm

- **1. That a single number can “best” represents an entire data set is known as**
- a) Covariance
- b) The Intercept
- c) **Measure of Central Tendency**
- d) The concept of “least squares”

Sample Questions for Midterm

- **2. Which of the following is most closely associated with dispersion**
- **a) Standard deviation**
- b) Non-normal distributions
- c) Confirmation bias
- d) Slope

Sample Questions for Midterm

- **9. In a normal distribution, compared to the median, the mean will be**
- **a) The same**
- b) One standard deviation higher
- c) Always less than the median
- d) Always greater than the median

Sample Questions for Midterm

- **10. A benefit of the mean absolute deviation is**
- a) It can still be used in calculating deviations from the average
- b) It removes the problem where subtracting observations from the mean (of a data set) sums to zero
- c) As compared to the standard deviation, it is impacted less by outliers
- **d) All of the above**

Probability Theory

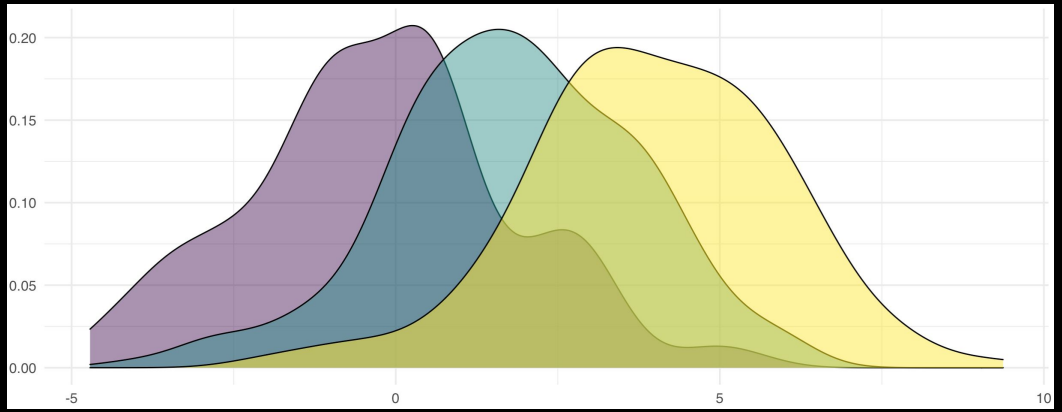
Review:

→Event Space, Probability Space

Conditional Probability

Frequentist Inference

Bayesian Inference



NYU

Sample Questions for Midterm

- **7.1 With two independent events, the likelihood (probability) of both occurring is**
- a) 1.0
- b) The probability of the first event plus the probability of the second event
- **c) The probability of the first event multiplied by the probability of the second event**
- d) 0

Sample Questions for Midterm

- **7.2 With two mutually exclusive events, the likelihood (probability) of both occurring is**
- a) 1.0
- b) The probability of the first event plus the probability of the second event
- c) The probability of the first event multiplied by the probability of the second event
- **d) 0**

Sample Questions for Midterm

- **7.3 With two mutually exclusive events, the likelihood (probability) that the two events are independent is**

  a) 1.0

- b) **0**
- c)The probability of the first event divided by the probability of the second event
- d) cannot be defined from given information

Sample Questions for Midterm

Probability can be used to make inference about the rough relationship between variables(events). Which of the following is true given two variables A and B?

a) $P(A,B)=P(A)+P(B)$
b) $P(AB) = P(A)P(B)$
c) $P(A)+P(B) = 1$
d) $P(AB) = P(A,B)-P(A)-P(B)$

Sample Questions for Midterm

If we can observe more events(more than 2), Which of the following is true given two variables A,B and C

a) $P(A,B,C)=P(A)+P(B)+P(C) - 2P(ABC)$
b) $P(ABC) = P(A)P(B)P(C)$
c) $P(A)+P(B)+P(C)<= 1$
d) P(not A, not B, not C)= 1- P(A,B,C)

Probability can be complex but do not forget to use **Venn Diagram**

Practice For Probability Theory

See notes Notes for Probability theory

Probability Theory

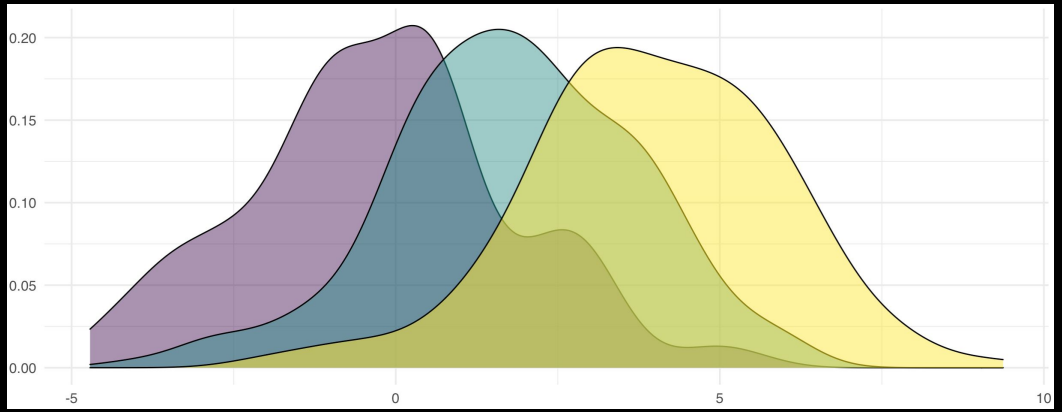
Review:

Event Space, Probability Space

→Conditional Probability

Frequentist Inference

Bayesian Inference



NYU

Sample Questions for Midterm

- **The New York University just conducted a short survey to collect the data if students feel satisfied with the remote-courses. preliminary academic studies have assigned the following probabilities of finding satisfaction**

  P (high satisfaction) = a

  P (medium satisfaction) = b

  P (no satisfaction) = c

- **Does the sum of three probabilities to 1?**

Sample Questions for Midterm

- **The New York University just conducted a short survey to collect the data if students feel satisfied with the remote-courses. preliminary academic studies have assigned the following probabilities of finding satisfaction**

  P (high satisfaction) = a

  P (medium satisfaction) = b

  P (no satisfaction) = c

- **Does the sum of three probabilities to 1?**

Yes, because there are only three levels of satisfaction.

The sum of the probability are mutually exclusive so the sum of three represent the whole world of satisfaction in this study

Sample Questions for Midterm

- **After 200 courses enrollment of the first semester, an enrollment test is taken. The probabilities of finding the particular type of enrollment identified by the test are as follows:**
- P (enrollment | high satisfaction) = 0.20
- P (enrollment | medium satisfaction) = 0.80
- P (enrollment | no satisfaction) = 0.20

Justify the answer for probability of enrollment of courses

Should P(enrollment) be 1 ?

Sample Questions for Midterm

- **After 200 courses enrollment of the first semester, an enrollment test is taken. The probabilities of finding the particular type of enrollment identified by the test are as follows:**
- P (enrollment | high satisfaction) = 0.20
- P (enrollment | medium satisfaction) = 0.80
- P (enrollment | no satisfaction) = 0.20

Justify the answer for probability of enrollment of courses

Should P(enrollment) be 1 ?

No necessarily to be true, see why in the whiteboard notes

Practice for conditional probability

See note

Probability Theory

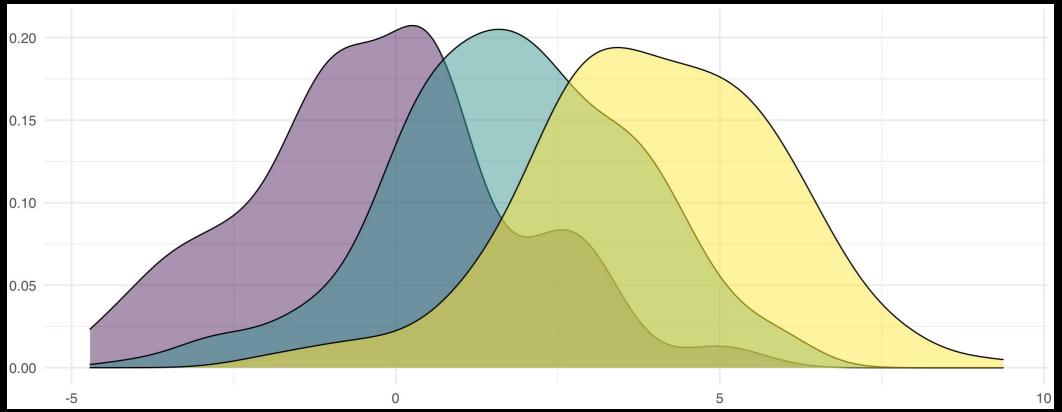
Review:

Event Space, Probability Space

Conditional Probability

→Frequentist Inference

Bayesian Inference



NYU

Sample Questions for Midterm

11. **In researching the chances of getting into a particular graduate school, a student decides to look at past acceptance rates. Using this information would be an example of**

a) Confirmation bias

b) Causality : *causality need fair comparison*

c) Inductive reasoning

d) Frequentist statistics : *objective evidence based on experiments and historical data*

Note on Frequentist Inference

There are two major schools of thought for statistics: frequentist inference and Bayesian Inference.

It interprets probability as the long term frequency. In Frequentist approach, the parameter of interest **is a fixed and unknown number.**

To be specific, we are interested in what one or a set of parameter(s) will better generate and represent our observations.

The past data is our observations and can be used to make inference about the population.

Sample Questions for Midterm

4. Confirmation bias indicates

a) That people are suspicious of a sequence of positive outcomes (e.g. 5 “heads” in a row in a series of coin tosses)

b) The greater likelihood of seeking supporting evidence for a claim than contradicting evidence

c) A perceived correlation that is only the result of one or more outlier values of a data set

d) A general trend wherein in most of a sample is likely to appear one standard deviation away from the mean (that is one standard deviation above or below the mean)

Sample Questions for Midterm

5. In some cases of deductive reasoning (for example with the Wason Card Selection Task), a full testing of the associated reasoning lacks

a) Dispersion

b) Correlation

c) Falsification

d) Subjective modeling

Could you find some classic example of falsification?

Sample Questions for Midterm

Could you find some classic example of falsification?

A business school claims that they have a very strict and competitive selection process when they admit prospective students into their MBA programs. According to them, the average GMAT score of the admitted students is at least 720. As a journalist who is concerned about the fairness of the admission process of this school, you suspect that the school is falsifying such claims in order to increase the tuition fees.

In this study, you observe there was a negative relationship between Tuition fee and GMAT score because the school propagandize their average admitted students' score is high than 720.

Probability Theory

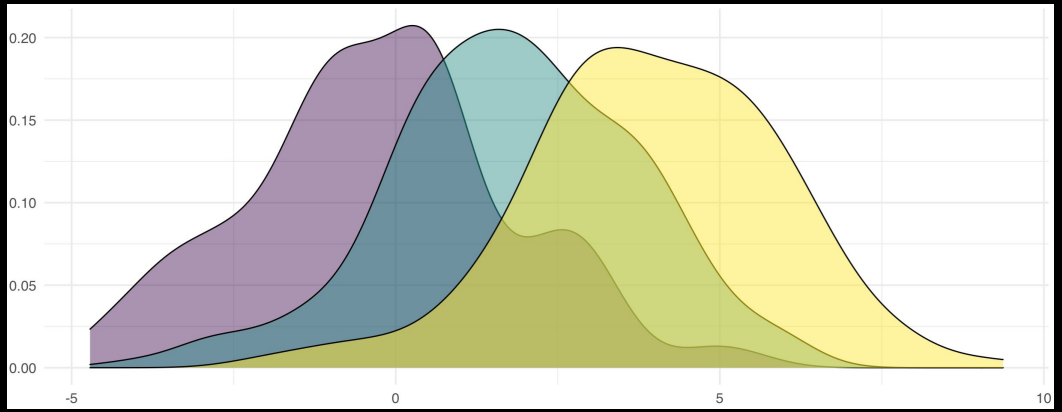
Review:

Event Space, Probability Space

Conditional Probability

Frequentist Inference

→Bayesian Inference



NYU

Note on Bayesian Inference

The Bayesian inference is an alternative statistical paradigm to the Frequentist approach. The Bayesian approach interprets the probability in a broader sense that include subjective probability, which allows us to assign probability to almost every quantity in our model (including the parameter of interest and even a statistical model).

The Bayesian inference relies on a simple decision theoretic rule – if we are competing two or more choices, **we always choose the one with higher probability.** This simple rule allows us to design an estimator, construct an interval, and perform hypothesis test.

Note on Bayesian Inference

The Bayesian inference is an alternative statistical paradigm to the Frequentist approach.

Bayesian inference set a an original belief: for example you assume the **prior probability** of flipping a coin and getting a head is 0.5.

Then, you get more times of results and you regard the results as additional information and **use the additional information to update your original belief.**

The updated probability is called posterior probability

The second posterior probability is then used as new prior to update the third time posterior and you repeat the process and closer to the true.

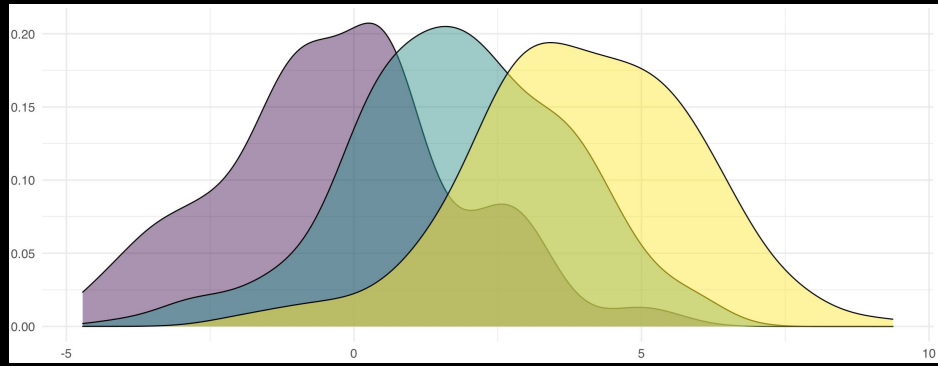
Whiteboard Practice

See notes Cov/cor/var/reg

Correlation, slope and linear regression

Review:

→Correlation, Simple Linear Regression



NYU

Sample Questions for Midterm

- **2.A Pearson's r analysis of 1.0 can be described as**
- **a) Perfect, but still lacking causality**
- b) Random
- c) Impossible
- d) An inverse relationship between the two variables being analyzed

Sample Questions for Midterm

Overall, as an association between two variables increases, covariance would be expected

a**) Gets larger**

b) Gets smaller

c) The covariance value is independent of an association between two variables

d) Can only be greater than 0

Sample Questions for Midterm

Linear correlation assumes that two variables, when plotted together,

a) Roughly parallel the x-axis

b) Form a line whose intercept is 0 and whose slope is 1

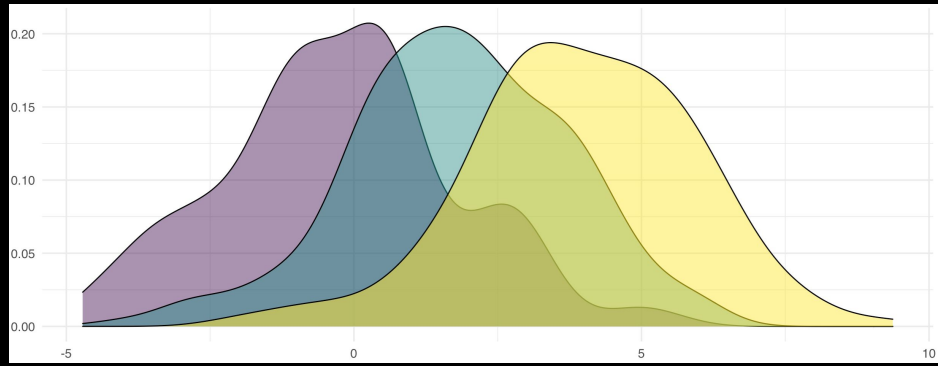
c) Are best described by a single line

d) Have half of the plotted points showing positive values and the other half negative

Correlation, slope and linear regression

Review:

→Correlation, non-linear correlation



NYU

Note on Non-linear correlation

Good Luck on your Exam 1

NYU