# Data Types and Common Tools

Donna Dietz

American University

*dietz@american.edu*

STAT 202 - Spring 2020

# Data Types

When we deal with data, we need to be keenly aware of what our data "types" are.

- Numerical Data
  - Integers
  - Continuous Values (represented by Decmials)
- Non-numerical Data (Words)
  - Categorical (or "bins") without natural orderings
  - Ordinal (Categorical with a natural sequence)

# Data Types

Let's categorize some examples together:

- Flavors of ice cream
- Gallons of water in several local area pools
- School types, where schools can be preschool, high school, elementary etc.
- Frequencies a singer can sing
- Notes which can be played on a piano

# Overlapping Data Types!

Did that last one cause some confusion!?

It should!

Often your data will be classified by you for your purposes when you are the one analysing the data.

A singer's notes can be represented by continuous numerical data, by freqency.

However, for the piano, you could make good arguments for it being ordinal, categorical, integer, or continuous data! Think about why this might be the case!

# Answers

- Flavors of ice cream (categorical)
- Gallons of water in several local area pools (continuous)
- School types (ordinal)
- Frequencies a singer can sing (continuous)
- Notes which can be played on a piano
  - categorical: if you want to classify them by note name "A", "C" etc.
  - continuous: if you use the frequency, especially if it's in the process of tuning or going out of tune
  - integer: if you are transcribing music onto a staff or MIDI file
  - ordinal: if you realize the freqencies aren't equally spaced without a log scale and you just want to make sure the order is maintained

# Common Tools for Exploring Relationships

Now that you know it's important to figure out what your data types are, let's do something useful with that knowledge.

In your real-world life, after college, you may be asked to figure out if someone's techniques or methods were meaningful. The heart of what we do in data analysis is to determine whether there are relationships between variables or not, and if so, what those relationships are. These three tools are used very frequently, at least as a first-pass at analysis of data. You classify the tools you can use on your data by what your data types are.

# Common Tools for Exploring Relationships

## LSR

Least Squares Regression (LSR) is a tool you can use when you have at least two sets of numerical data, and you suspect the relationship may be simple. So, this is numerical vs. numerical. An example of a research question is: "Do children grow taller as they age?" You can measure age in years (or months) and height in inches (or meters). Both are numerical.

# Common Tools for Exploring Relationships

## ANOVA

Analysis of Variance (ANOVA) is a tool you can use when you think there might be differences between groups (categories) but the trait you are measuring is numerical. So, this is numerical vs. category. An example of a research question is: "Are men typically taller than women?". You have two groups of people, and in each group you can take a numerical measurement (height). Since you have exactly two groups, you could also use a "T-test".

# Common Tools for Exploring Relationships

## Chi-Square

Chi-Square ($\chi^2$) is a tool you can use when you are concerned about disproportionality of traits over various groups. Your data are usually represented as counts in a Two-Way table. So, this is category vs. category. An example of a research question is: "Are 8th grade boys or 8th grade girls more likey to know how to swim in deep water?". You have two groups of people again, and in each case you have a yes/no (categorical) question to ask each person. You want to know if the ratios of yes/no are the same for the boys as for the girls. Since you are comparing categories against other categories, you can use a Chi-Square test to determine if the ratios are different or the same.

# Try it!

Try classifying a few research questions and suggesting a good tool to use:

- You want to know if AU students spend more time in the gym during the Spring versus Fall semesters.
- You want to know if class year in college effects stated career goals.
- You want to know if students who spend more time watching videos spend more or less time socializing with friends.

# Try it!

- You want to know if AU students spend more time in the gym during the Spring versus Fall semesters. (numerical/categorical: ANOVA)
- You want to know if class year in college affects stated career goals. (categorical/categorical: Chi-Square)
- You want to know if students who spend more time watching videos spend more or less time socializing with friends. (numerical/numerical: LSR)

# Looking ahead

One concept that plays a crucial role in discussions of data analysis is that of P-values. There is a cute song you should learn to the tune of "Row Row Row Your Boat".

https://www.causeweb.org/cause/resources/fun/songs/what-p-value-means

It is key to know, what P-value means.
It's the chance with the null you obtain data that's...
At least that extreme.

# P-values

We will learn about P-values in detail later in the semester. For now, I'd like you to know this much:

- A P-value is a probability so...
- it has to have a value between 0 and 1
- The closer it is to zero, the more unusual (significant) it is
- For most situations we will consider a P-value to be insignificant if it is is 0.05 or higher.
- Many applications require an extremely low P-value to be considered significant.

# P-values

So, for example, a P-value of 0.01 may or may not be significant, depending on your research requirements. However, a P-value of 0.2 will never be considered significant, while a P-value of 0.00000000001 will nearly always be significant.

# Why do we care about probability?

Students often wonder what probability has to do with Statistics.

When we test our hunches (our hypotheses) we are thinking like this:

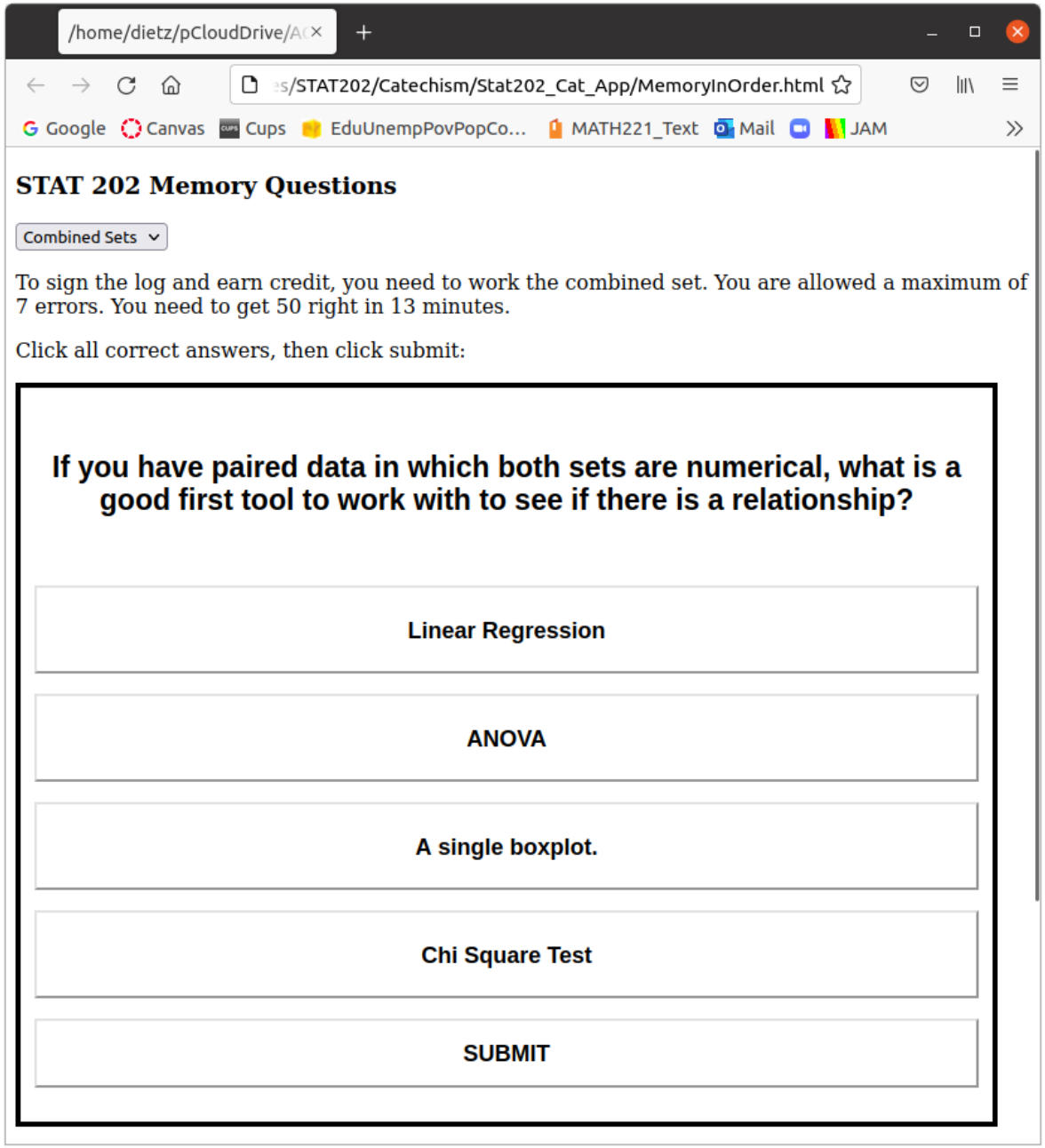
**Testing our guesses:**

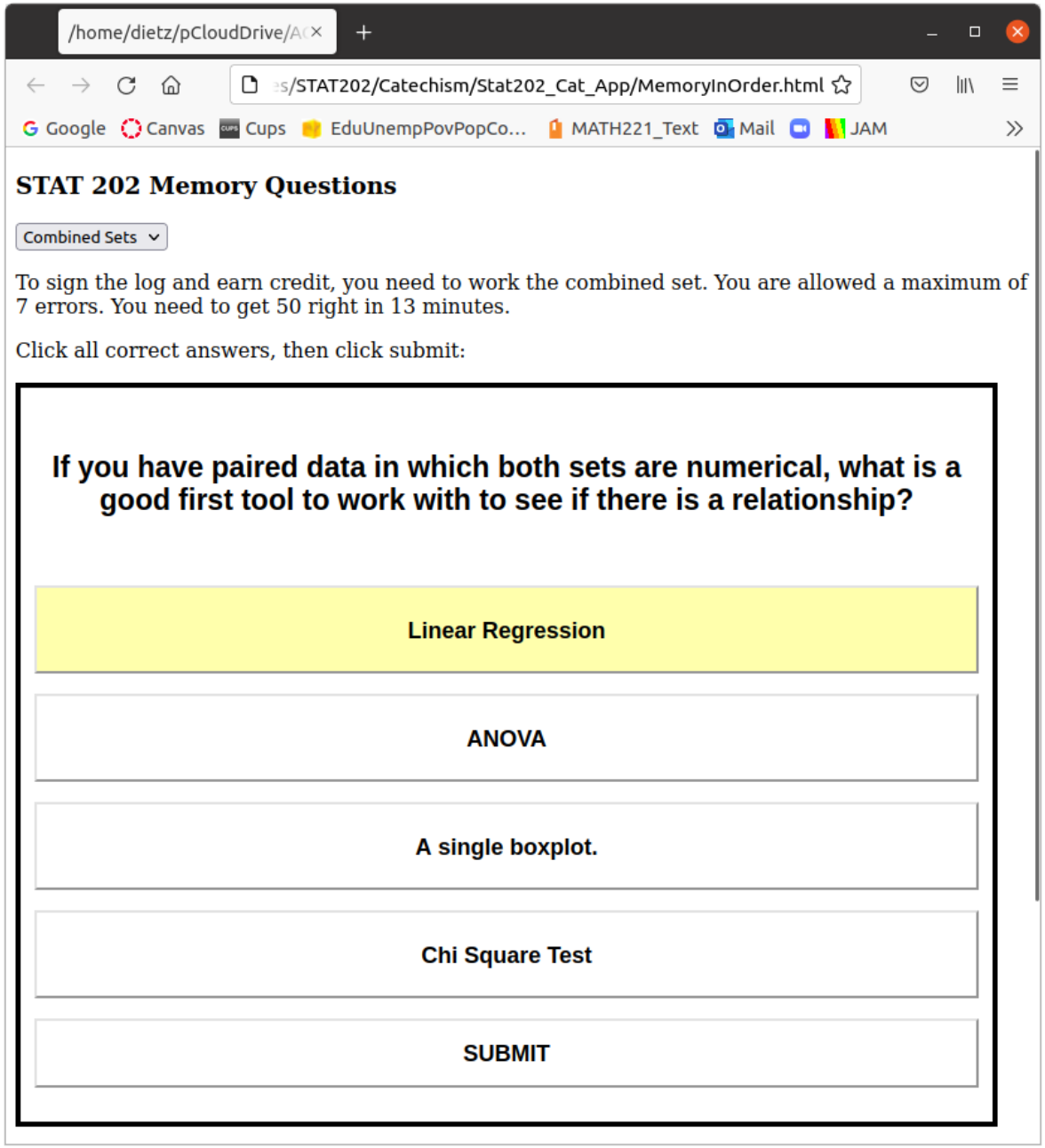
If we knew what our populations really looked like, we could estimate what a random sample of them would be likely to look like. Given that we have a random sample from a population, we'd like to know if our guesses are even possibly correct, or if they are likely to be close enough to our guesses.

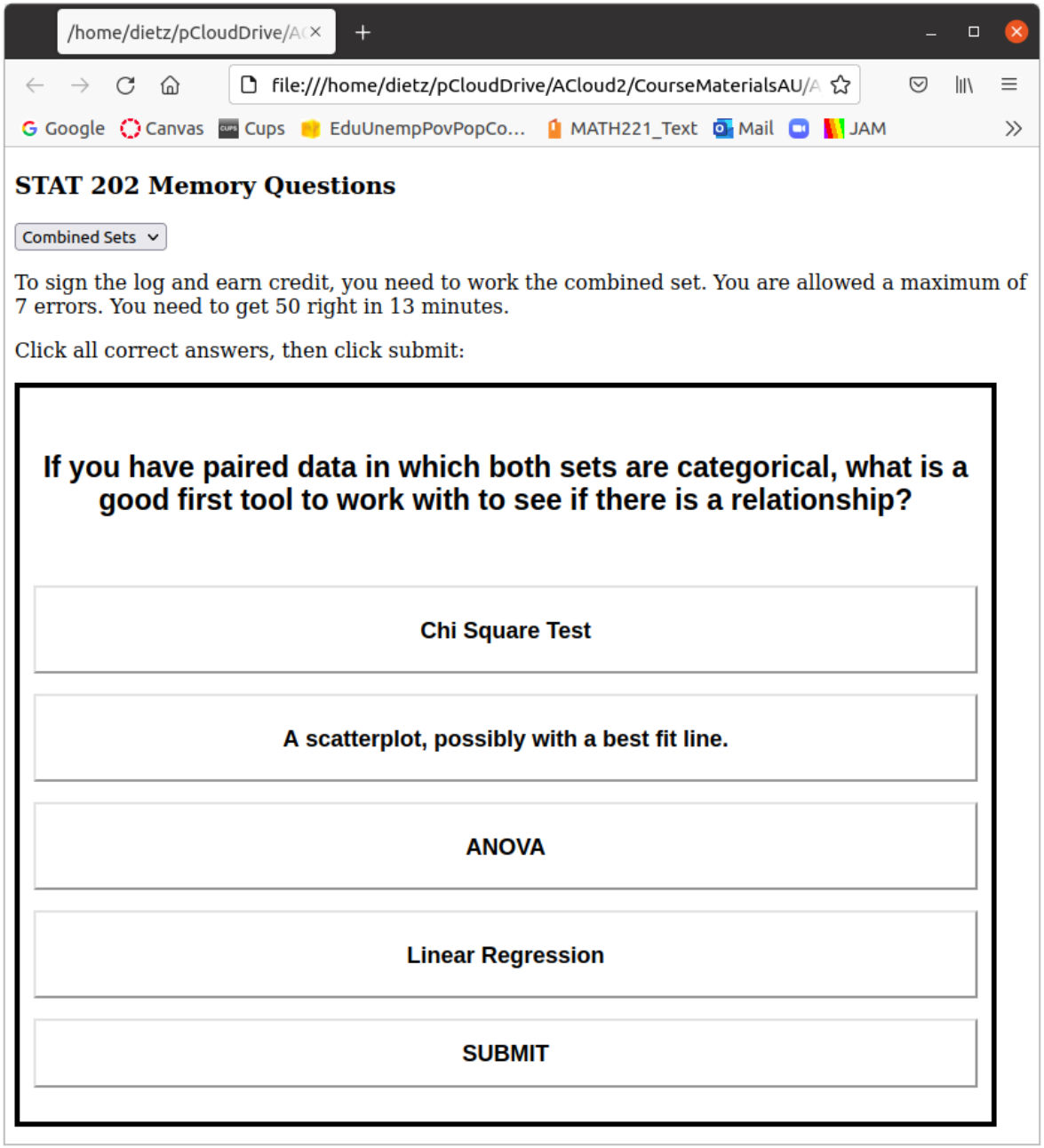
The rest of the course attempts to explain the core concepts of this simple idea.

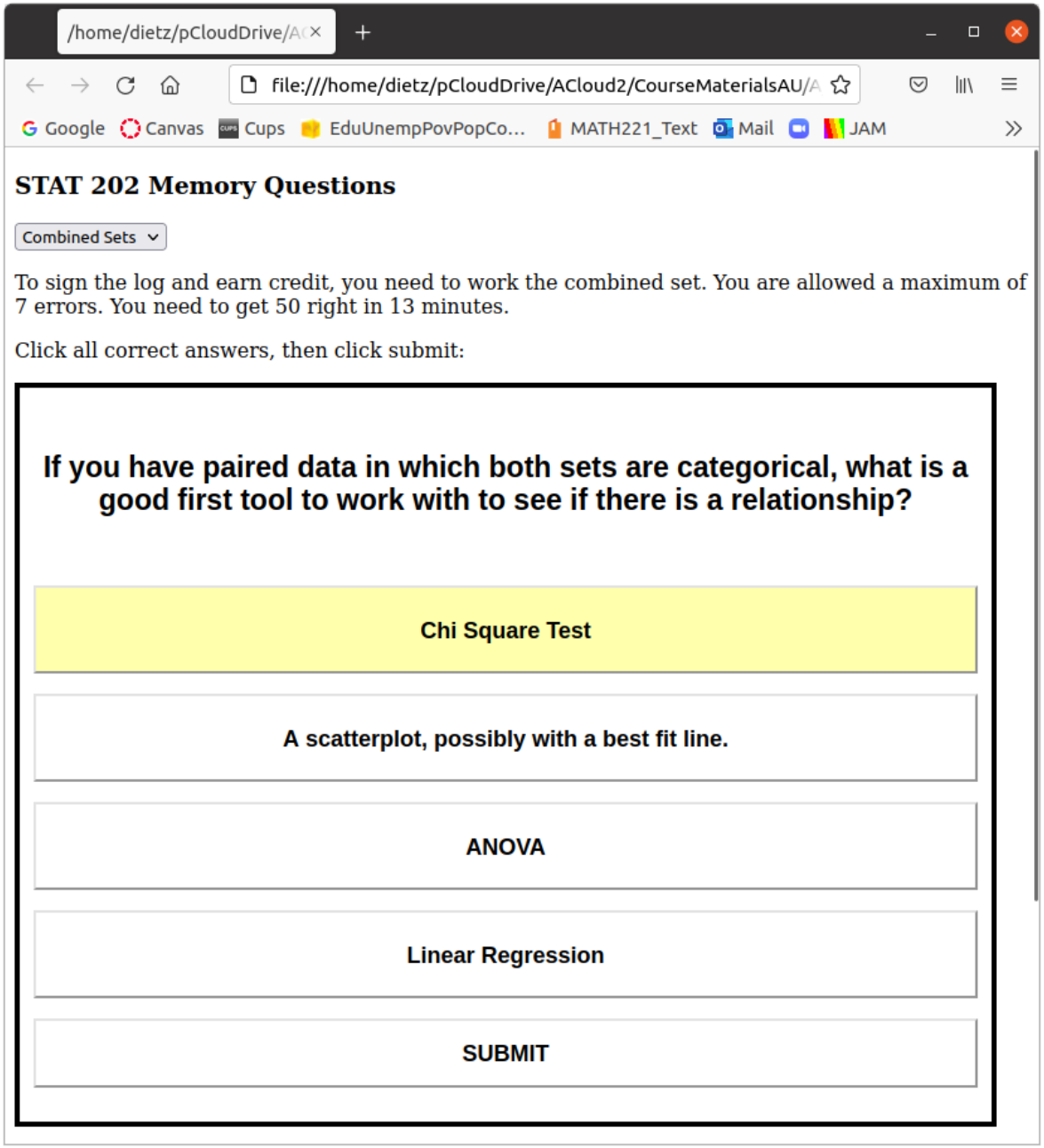
## MEMORY QUESTIONS

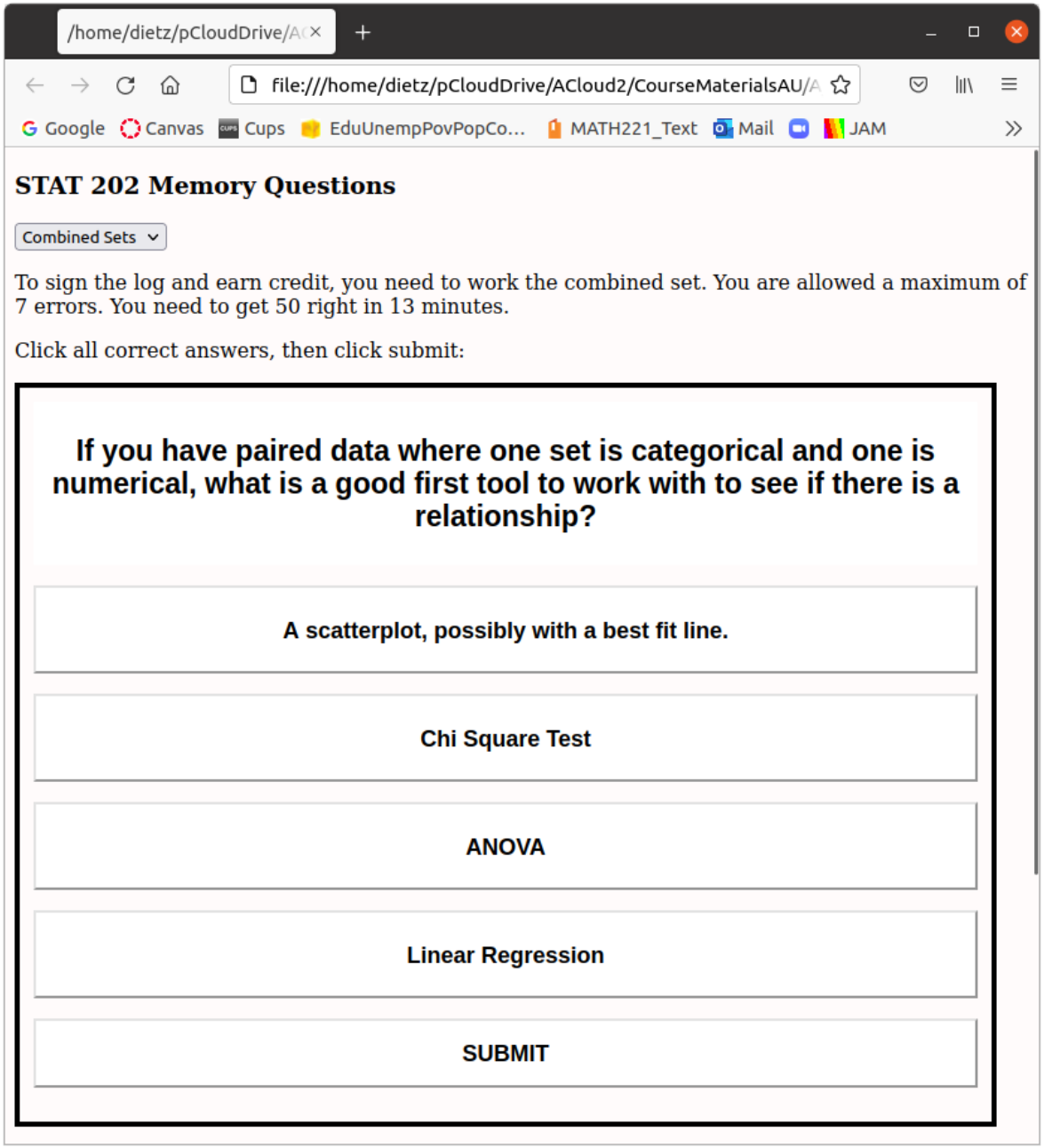
This semester, you will have about 100 memory questions. Half of them go with exam 1. Today we will go over the first 8.

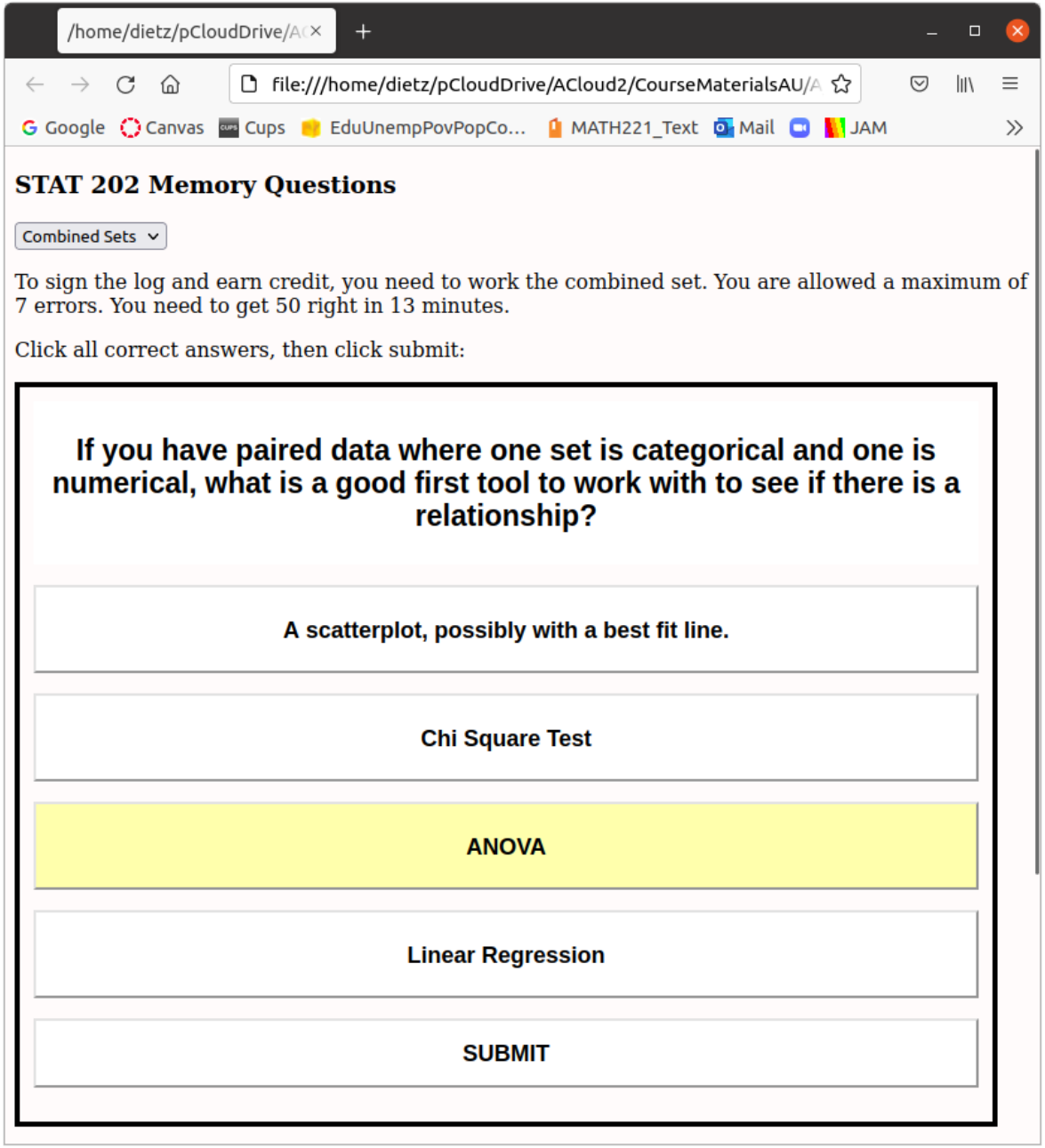

/home/dietz/pCloudDrive/A
s/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
If you have paired data in which both sets are numerical, what is a good first tool to work with to see if there is a relationship?
Linear Regression
ANOVA
A single boxplot.
Chi Square Test
SUBMIT

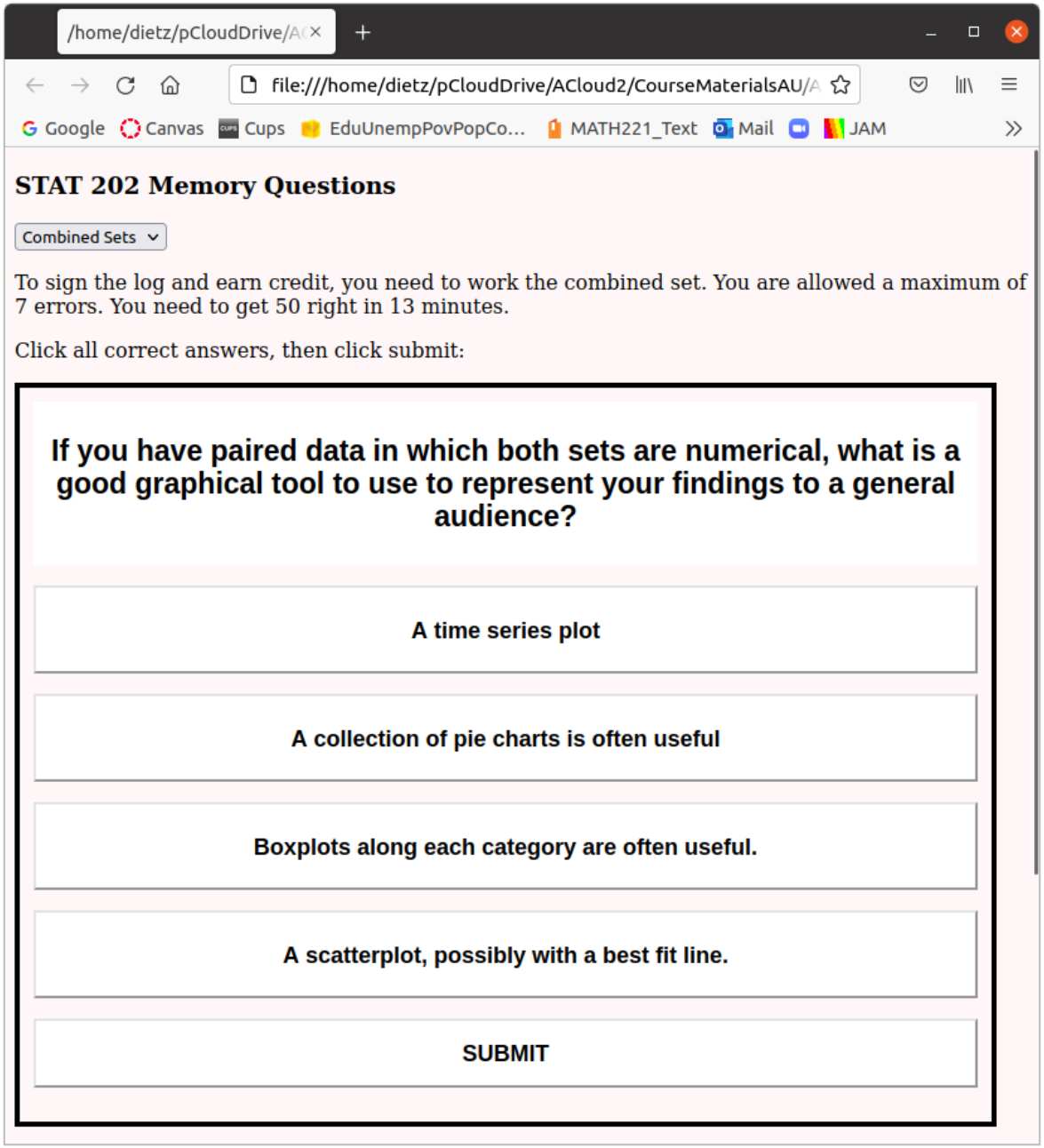

/home/dietz/pCloudDrive/A
s/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
If you have paired data in which both sets are numerical, what is a good first tool to work with to see if there is a relationship?
Linear Regression
ANOVA
A single boxplot.
Chi Square Test
SUBMIT

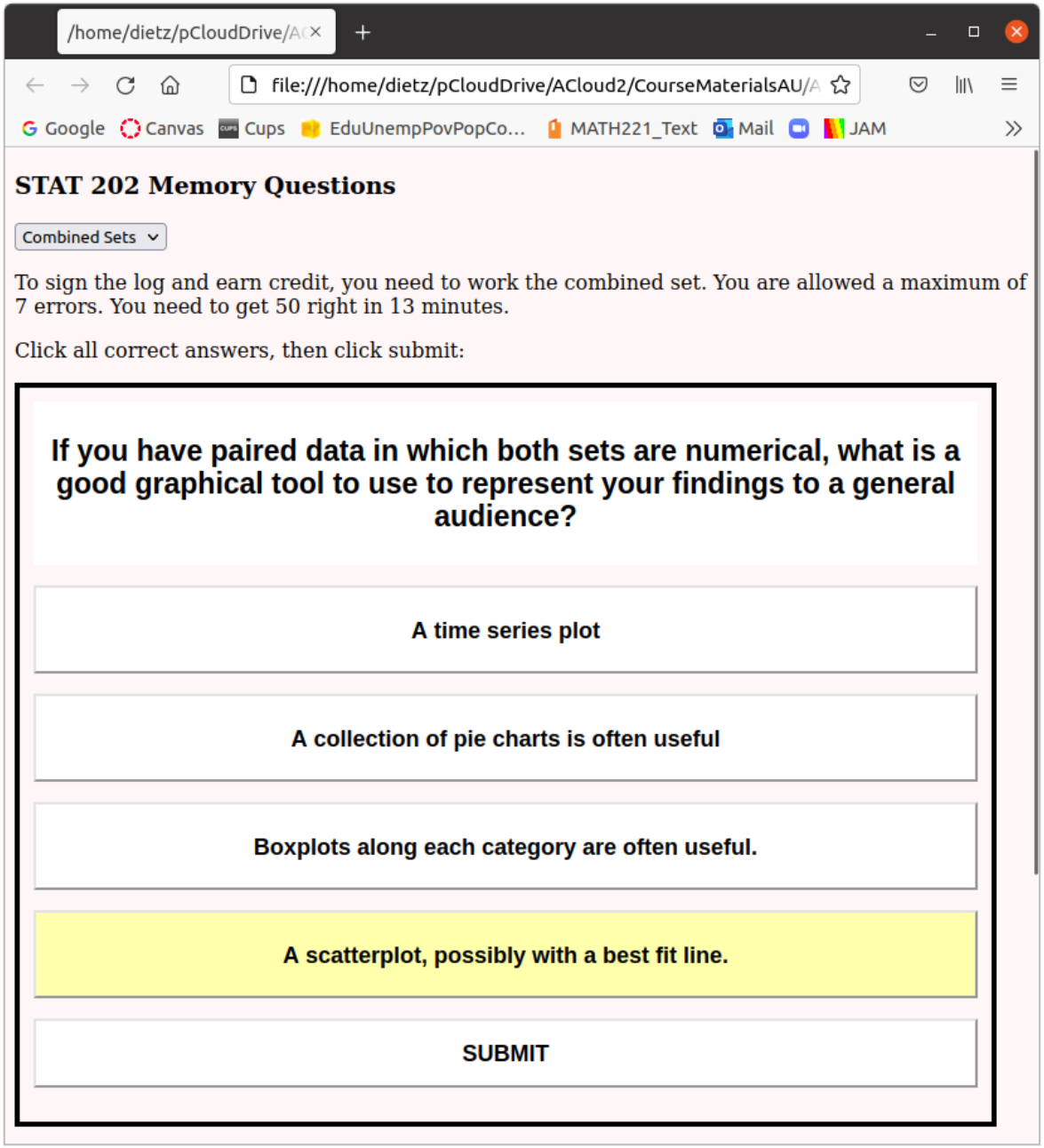

/home/dietz/pCloudDrive/A
file:///home/dietz/pCloudDrive/ACloud2/CourseMaterialsAU/A
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
If you have paired data in which both sets are categorical, what is a good first tool to work with to see if there is a relationship?
Chi Square Test
A scatterplot, possibly with a best fit line.
ANOVA
Linear Regression
SUBMIT

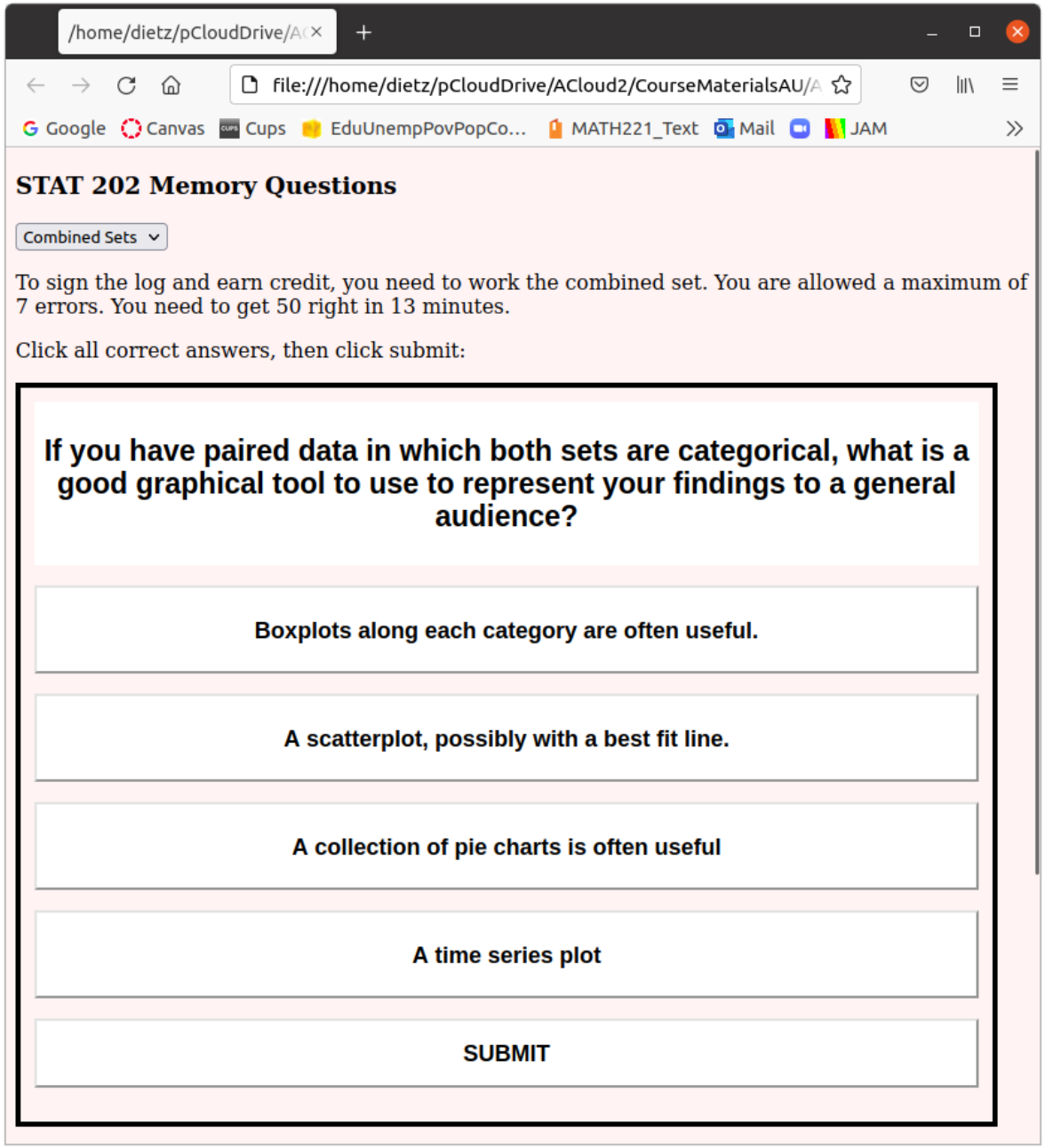

/home/dietz/pCloudDrive/A
file:///home/dietz/pCloudDrive/ACloud2/CourseMaterialsAU/A
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
If you have paired data in which both sets are categorical, what is a good first tool to work with to see if there is a relationship?
Chi Square Test
A scatterplot, possibly with a best fit line.
ANOVA
Linear Regression
SUBMIT

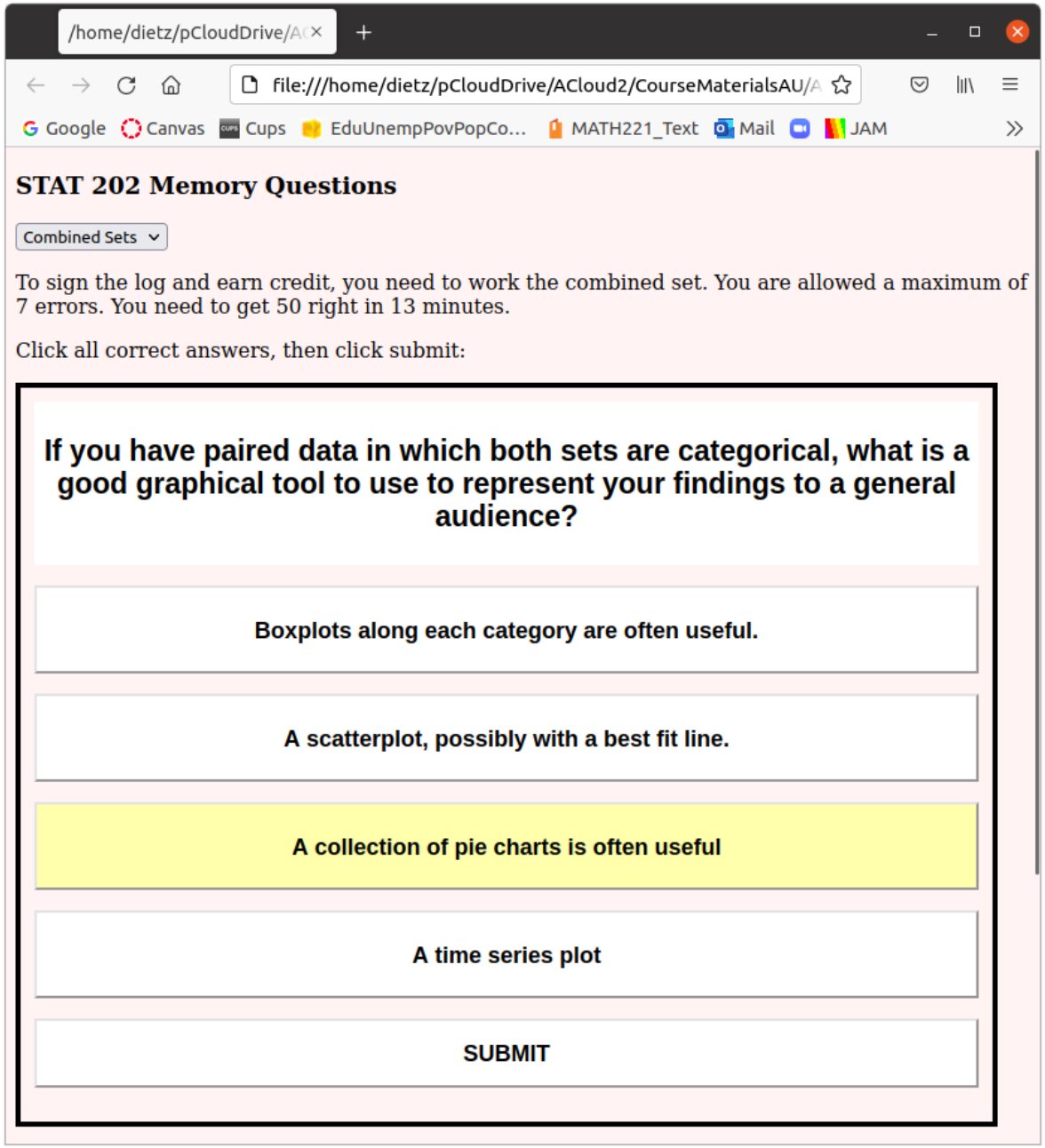

/home/dietz/pCloudDrive/A
file:///home/dietz/pCloudDrive/ACloud2/CourseMaterialsAU/A
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
If you have paired data where one set is categorical and one is numerical, what is a good first tool to work with to see if there is a relationship?
A scatterplot, possibly with a best fit line.
Chi Square Test
ANOVA
Linear Regression
SUBMIT

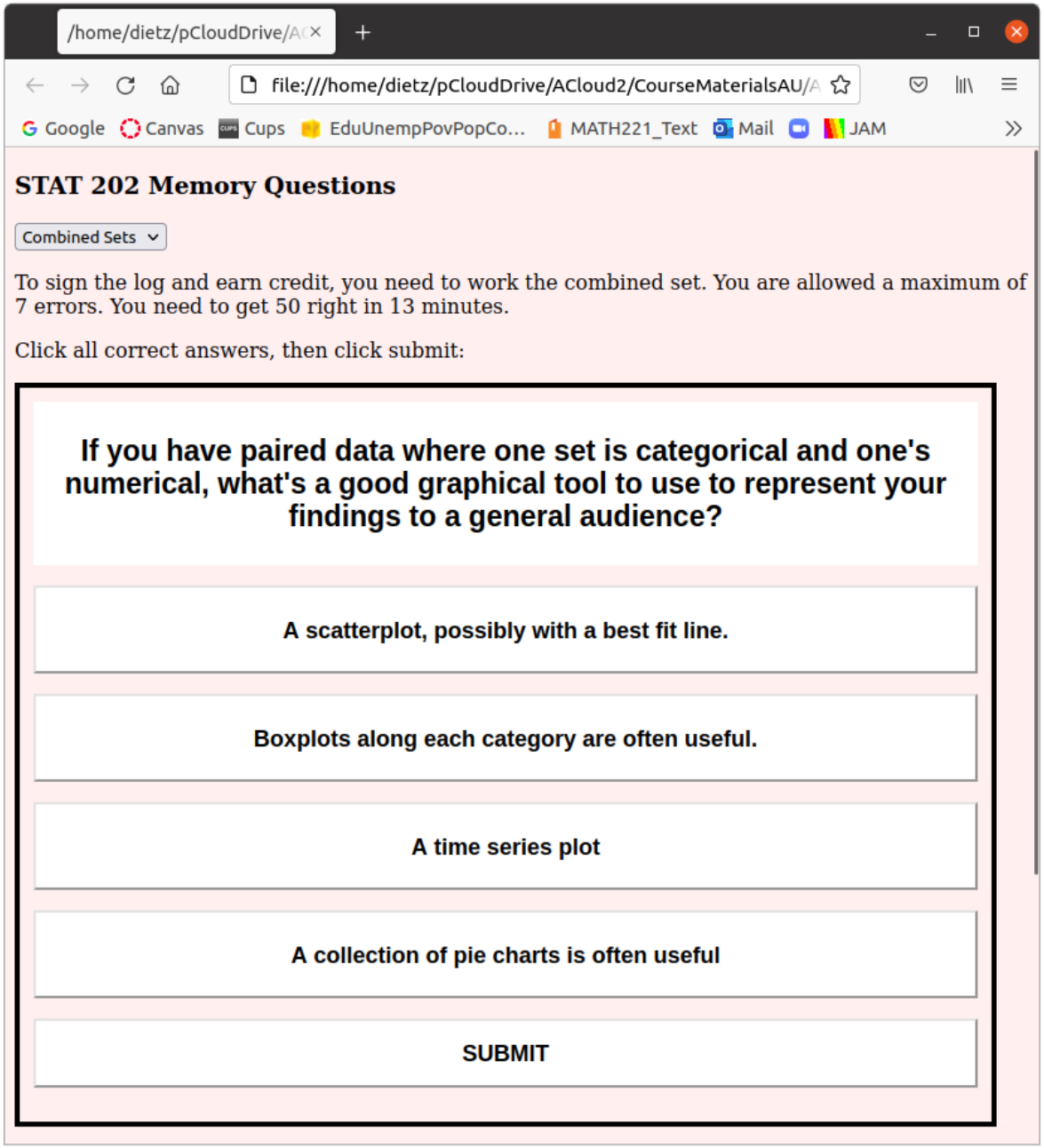

/home/dietz/pCloudDrive/A
file:///home/dietz/pCloudDrive/ACloud2/CourseMaterialsAU/A
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
If you have paired data where one set is categorical and one is numerical, what is a good first tool to work with to see if there is a relationship?
A scatterplot, possibly with a best fit line.
Chi Square Test
ANOVA
Linear Regression
SUBMIT

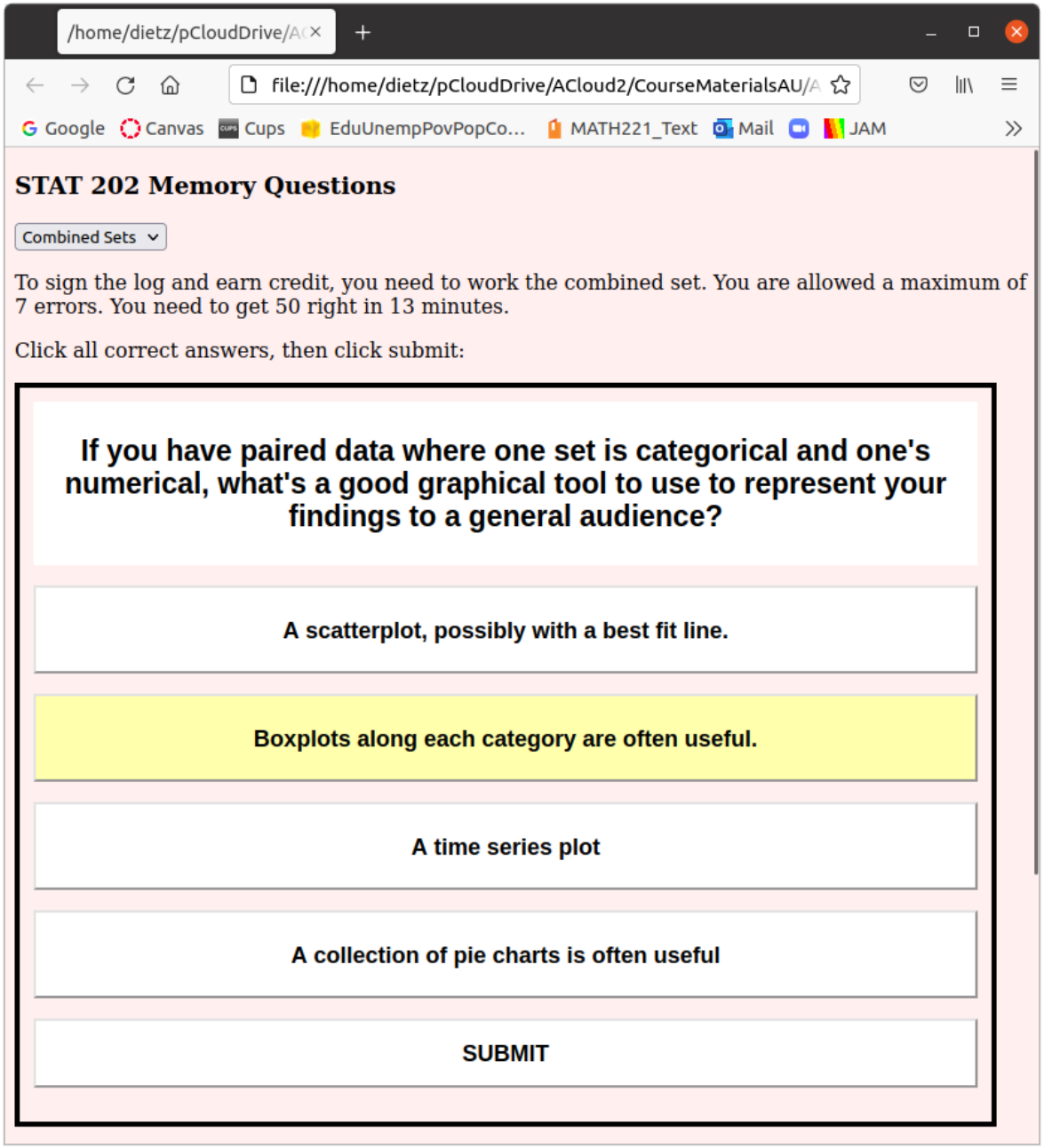

/home/dietz/pCloudDrive/A
file:///home/dietz/pCloudDrive/ACloud2/CourseMaterialsAU/A
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
If you have paired data in which both sets are numerical, what is a good graphical tool to use to represent your findings to a general audience?
A time series plot
A collection of pie charts is often useful
Boxplots along each category are often useful.
A scatterplot, possibly with a best fit line.
SUBMIT

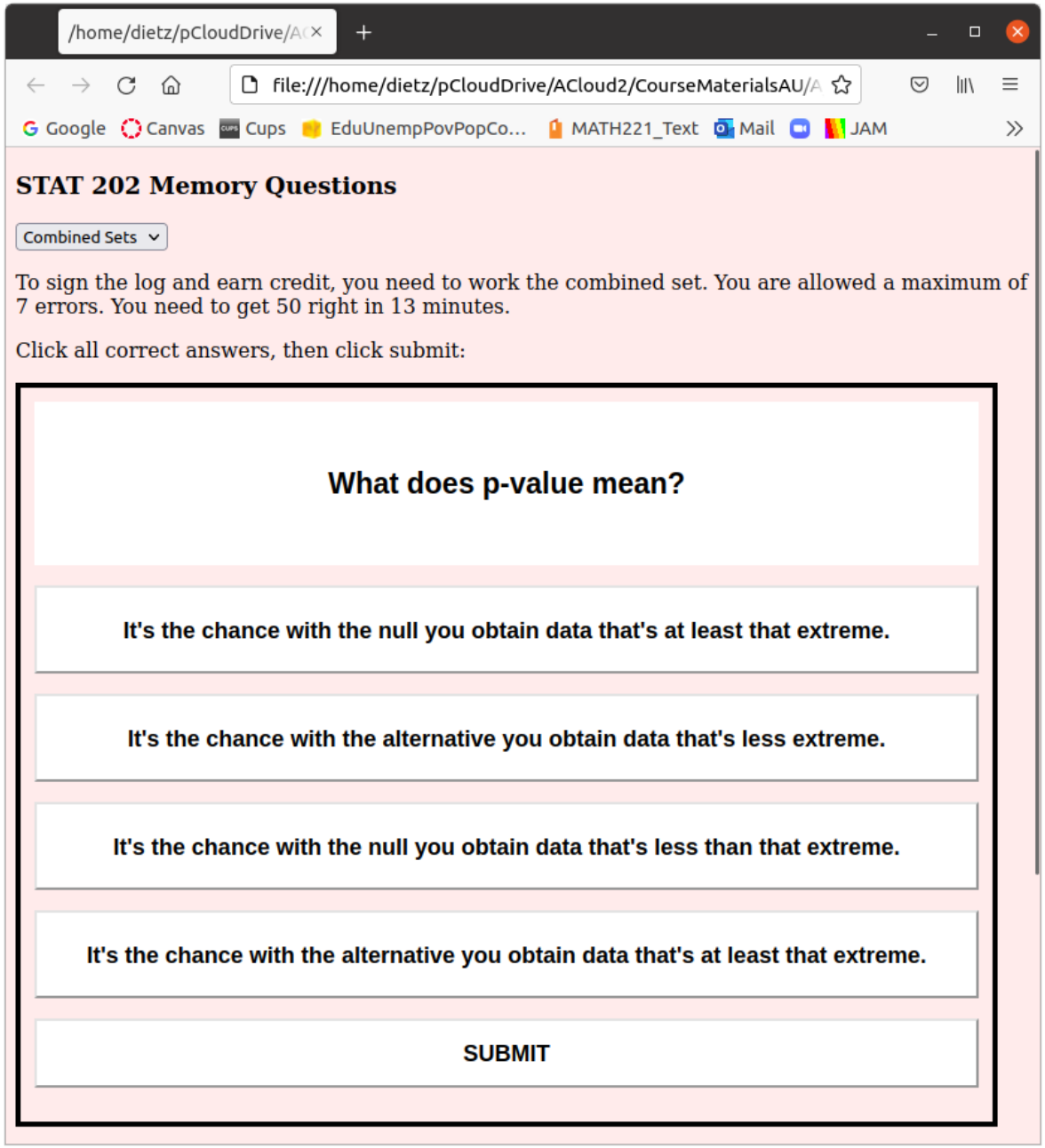

/home/dietz/pCloudDrive/A
file:///home/dietz/pCloudDrive/ACloud2/CourseMaterialsAU/A
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
If you have paired data in which both sets are numerical, what is a good graphical tool to use to represent your findings to a general audience?
A time series plot
A collection of pie charts is often useful
Boxplots along each category are often useful.
A scatterplot, possibly with a best fit line.
SUBMIT

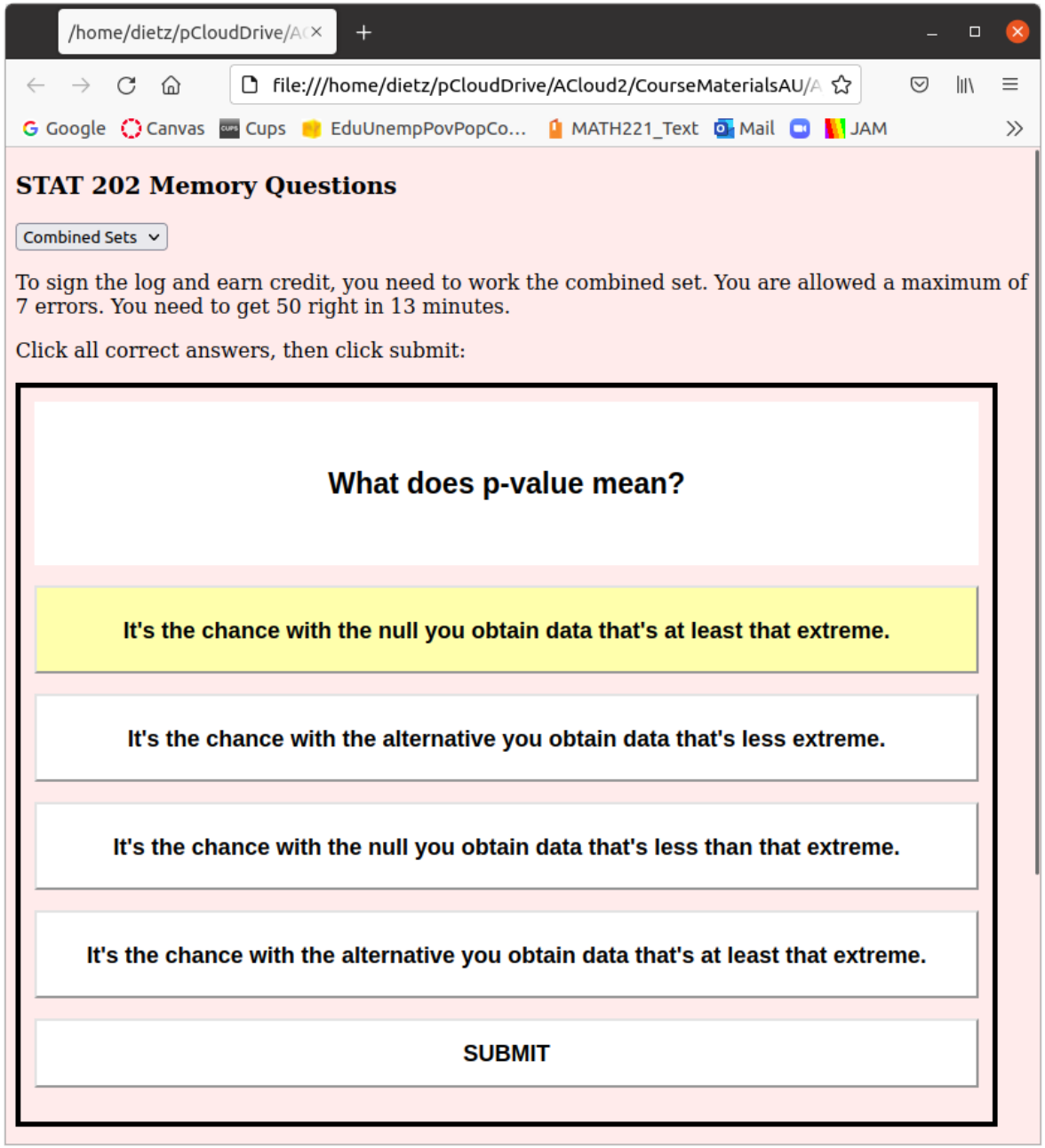

/home/dietz/pCloudDrive/A
file:///home/dietz/pCloudDrive/ACloud2/CourseMaterialsAU/A
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
If you have paired data in which both sets are categorical, what is a good graphical tool to use to represent your findings to a general audience?
Boxplots along each category are often useful.
A scatterplot, possibly with a best fit line.
A collection of pie charts is often useful
A time series plot
SUBMIT

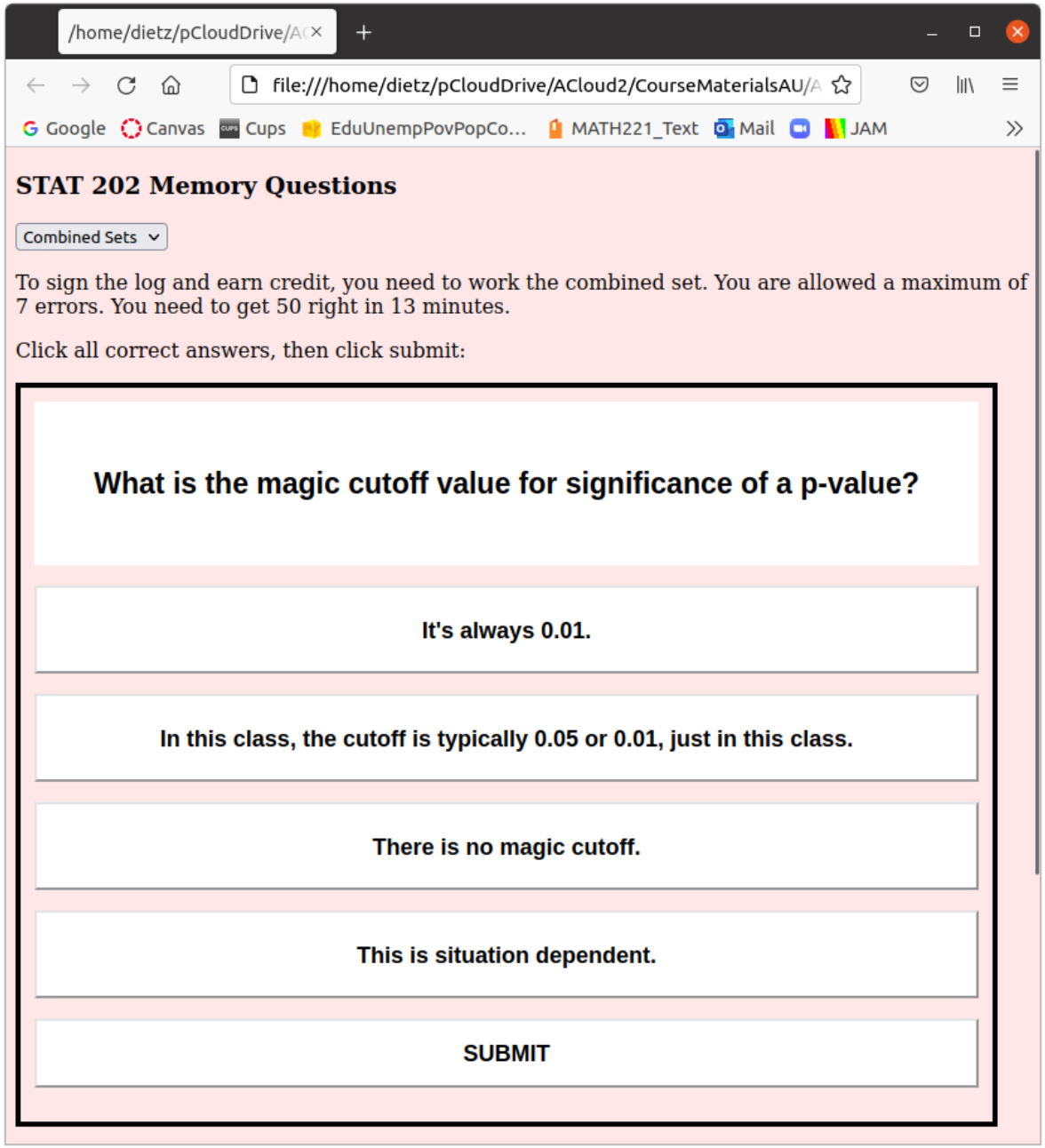

/home/dietz/pCloudDrive/A
file:///home/dietz/pCloudDrive/ACloud2/CourseMaterialsAU/A
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
If you have paired data in which both sets are categorical, what is a good graphical tool to use to represent your findings to a general audience?
Boxplots along each category are often useful.
A scatterplot, possibly with a best fit line.
A collection of pie charts is often useful
A time series plot
SUBMIT

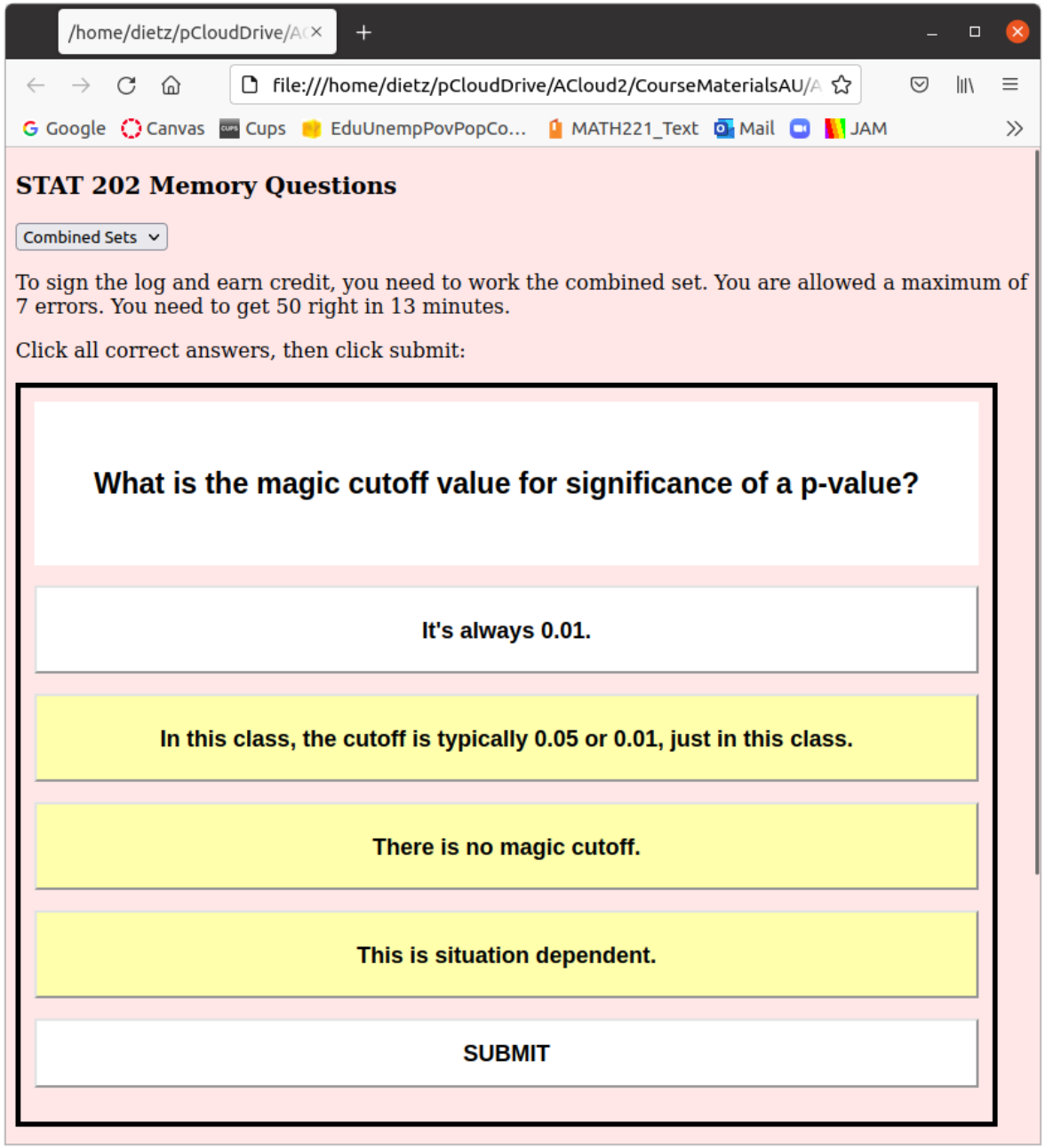

/home/dietz/pCloudDrive/A
file:///home/dietz/pCloudDrive/ACloud2/CourseMaterialsAU/A
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
If you have paired data where one set is categorical and one's numerical, what's a good graphical tool to use to represent your findings to a general audience?
A scatterplot, possibly with a best fit line.
Boxplots along each category are often useful.
A time series plot
A collection of pie charts is often useful
SUBMIT


/home/dietz/pCloudDrive/A
file:///home/dietz/pCloudDrive/ACloud2/CourseMaterialsAU/A
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
If you have paired data where one set is categorical and one's numerical, what's a good graphical tool to use to represent your findings to a general audience?
A scatterplot, possibly with a best fit line.
Boxplots along each category are often useful.
A time series plot
A collection of pie charts is often useful
SUBMIT


/home/dietz/pCloudDrive/A
file:///home/dietz/pCloudDrive/ACloud2/CourseMaterialsAU/A
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
What does p-value mean?
It's the chance with the null you obtain data that's at least that extreme.
It's the chance with the alternative you obtain data that's less extreme.
It's the chance with the null you obtain data that's less than that extreme.
It's the chance with the alternative you obtain data that's at least that extreme.
SUBMIT


/home/dietz/pCloudDrive/A
file:///home/dietz/pCloudDrive/ACloud2/CourseMaterialsAU/A
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
What does p-value mean?
It's the chance with the null you obtain data that's at least that extreme.
It's the chance with the alternative you obtain data that's less extreme.
It's the chance with the null you obtain data that's less than that extreme.
It's the chance with the alternative you obtain data that's at least that extreme.
SUBMIT


/home/dietz/pCloudDrive/A
file:///home/dietz/pCloudDrive/ACloud2/CourseMaterialsAU/A
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
What is the magic cutoff value for significance of a p-value?
It's always 0.01.
In this class, the cutoff is typically 0.05 or 0.01, just in this class.
There is no magic cutoff.
This is situation dependent.
SUBMIT


/home/dietz/pCloudDrive/A
file:///home/dietz/pCloudDrive/ACloud2/CourseMaterialsAU/A
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
What is the magic cutoff value for significance of a p-value?
It's always 0.01.
In this class, the cutoff is typically 0.05 or 0.01, just in this class.
There is no magic cutoff.
This is situation dependent.
SUBMIT