# Shapes of Distributions

Donna Dietz

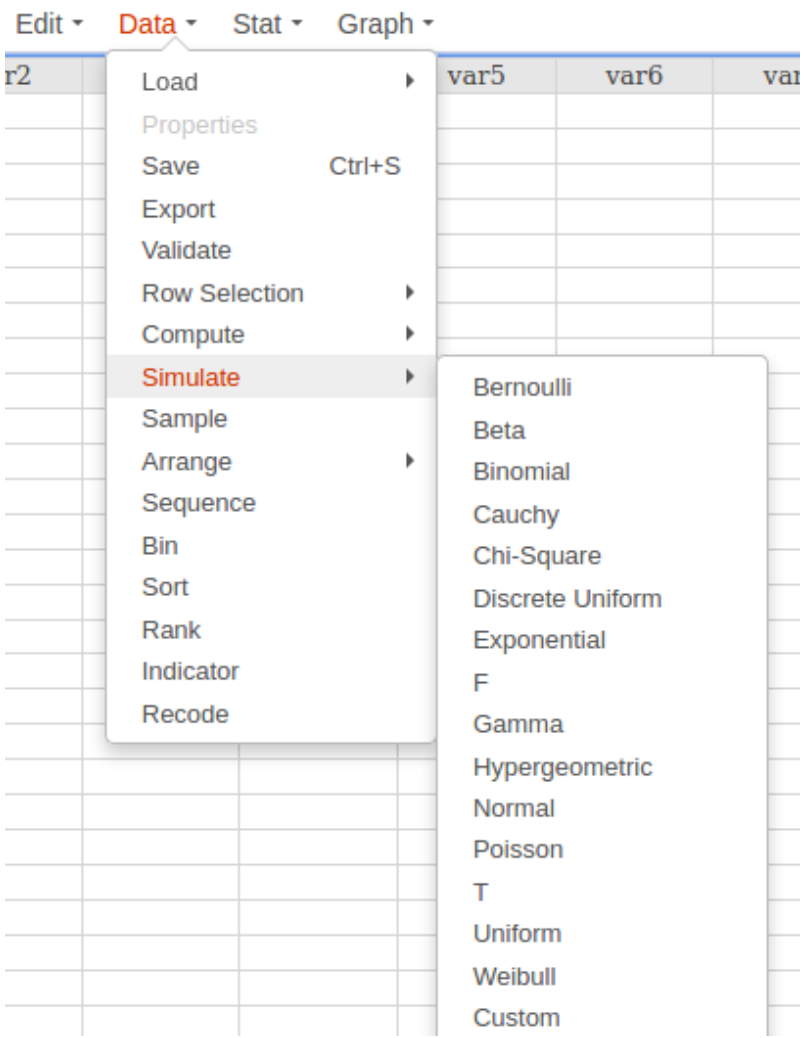
American University

*dietz@american.edu*

STAT 202 - Spring 2020

# Shapes of Distributions

Many types of distributions exist, each with a different shape. To give you an idea of the names of the more common ones, here are the ones you can easily simulate on StatCrunch:



# Shapes of Distributions

However, for all distributions, we want to have some common tools to describe them.

- Measures of Centrality
  - mean
  - median
  - mode
- Measures of Spread
  - IQR
  - Standard Deviation
- Symmetric, or Skewed?
  - Right Skewed
  - Symmetric
  - Left Skewed

# The Mean

The **mean** of a data set is the numerical average. This is usually the first measure of centrality students learn, so you probably already know this. This use of the Sigma (summation) symbol is common for Statisticians:

$$\bar{x} = \frac{\sum x}{n}$$

Mathematicians would write:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

# The Median

The **median** is the data point in the middle, if you line them up from smallest to largest. The median occupies the middlemost position in a lineup. For small data sets, if we have an even number of items, we take the average of the two middle-most items. For large sets, we probably won't have to use this hack, because many similar data points will lie close together.

# The Mode

The **mode** is the data point which occurs the most times. If you plot a histogram, it's often in the tallest bar. If you draw a continuous curve to represent your distribution, it's where the peak occurs. (Not all data have a mode, though!)

# Symmetry

What does it mean for a data set to be symmetric?

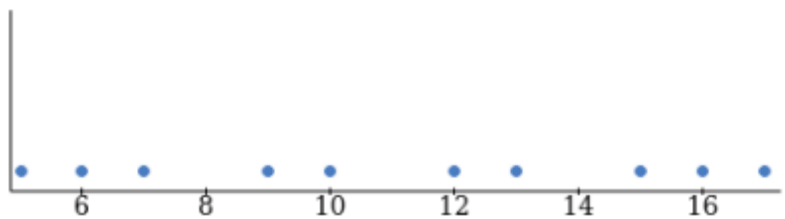
Think of plotting the data on a line. If there is a point in the middle that could act like a mirror (with or without a data point there), then that is the point of symmetry, and your data are symmetric!

# The Mean and Median with Symmetry

In words you can say:

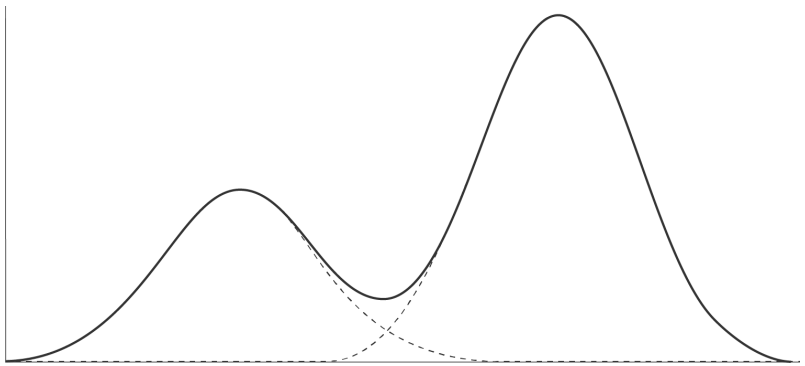
To find the mean of a set of data, add up all the data elements, then divide by the number of items in your set.

If you are lucky enough to have a symmetric data set, the mean and median will always be right in the middle, in the same spot!



# Data with more than one mode

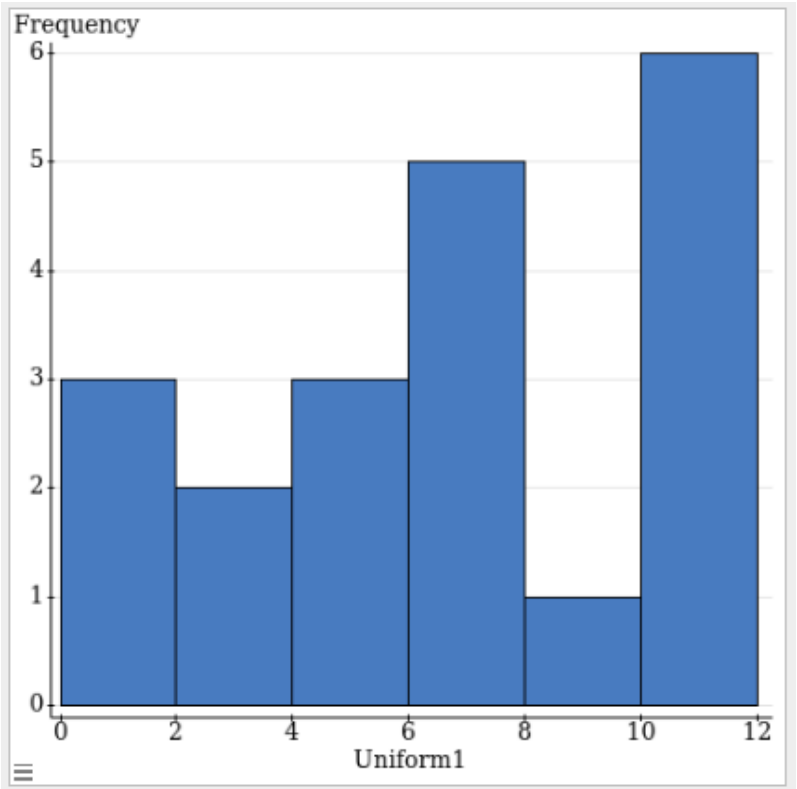
Although not all data have a mode at all, when they do, they could have two modes (bimodal) or many modes (multi-modal).



This graphic is from getnave.com and respresents this concept well. The purpose for declaring that a distribution has multiple modes is that you suspect you've really added two distributions together.

# Not Bimodal

This might be labeled as bimodal, because of the two peaks. Or, unimodal, because there is one actual maximum. But, I know this is just sampling error. The data were randomly drawn from a Uniform distribution, with no mode at all! Knowing where your data come from makes all the difference!



# IQR as a measure of spread

The inter-quartile range (or IQR) is a single value that represents the width occupied by the middle 50% of the data.

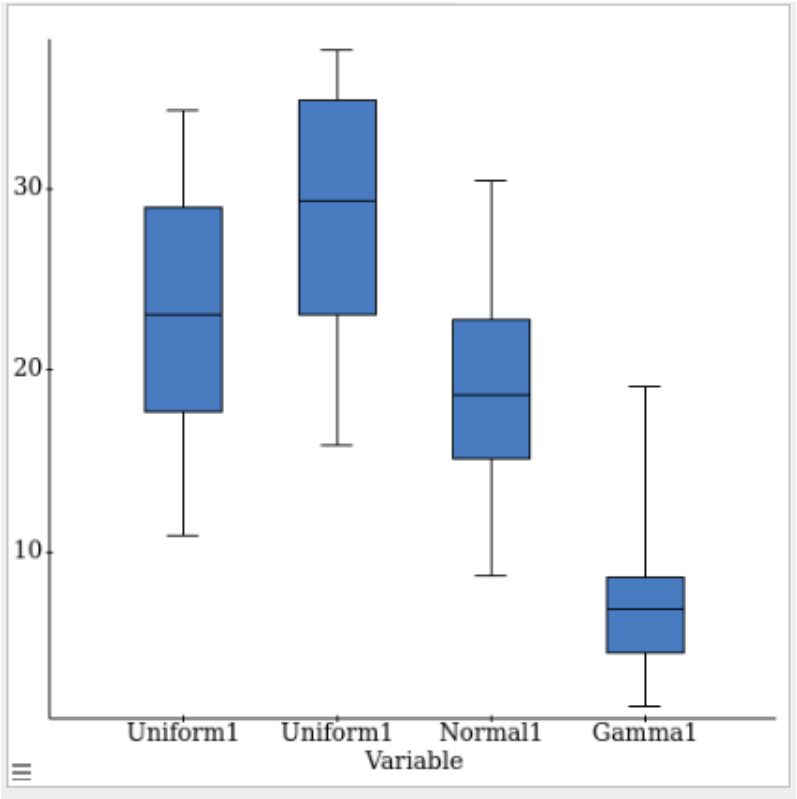
$$IQR = Q3 - Q1$$

# Five Number Summaries

The quartiles (Q1, Q2=median, Q3) together with the maximum and minimum of a data set is the **Five Number Summary**.

# Boxplots

Boxplots show visually where the five numbers of a five number summary lie. They are generally most useful when displayed with other boxplots.



# Outliers

Outliers are elements in a data set which are very far from the rest of the data. It's really hard to tell whether data are very far away from the rest of the data for a good reason, or if there was just a measurement error. Maybe something is contaminating your data set? Ultimately, you need to know where your data come from and why those values are in your set. You need to guard against the instinct to over-trim your data, but we do have rules of thumb.

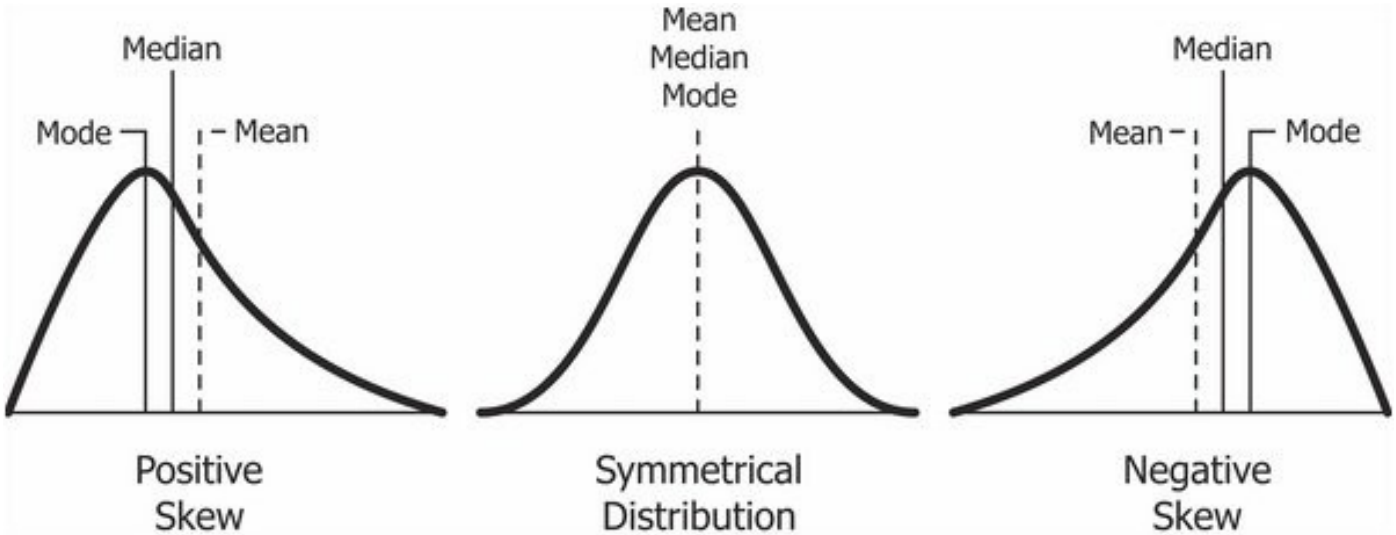
# Finding possible outliers

The rule often used to find potential outliers is any data outside this cutoffs:

$$\textit{lower cutoff} = Q1 - 1.5 \times IQR$$

$$\textit{upper cutoff} = Q3 + 1.5 \times IQR$$

# Skew




Positively skewed (Right-skewed) data sets have a longer tail on the upper end and are steeper on the left side. Negatively skewed (Left-skewed) have the reverse.

For Right-skewed data $mode < median < mean$.

For Left-skewed data $mean < median < mode$.

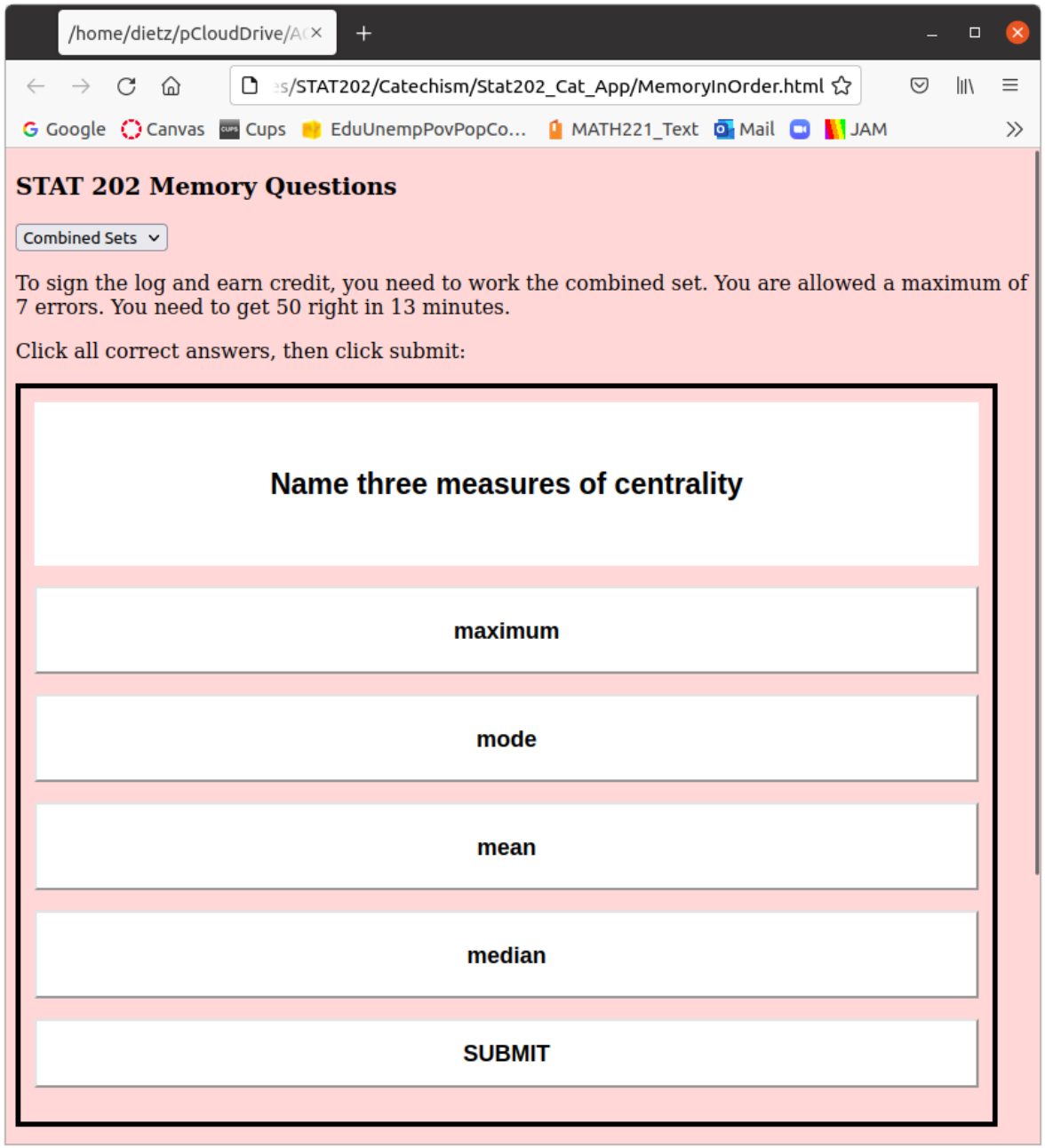
# Standard Deviation

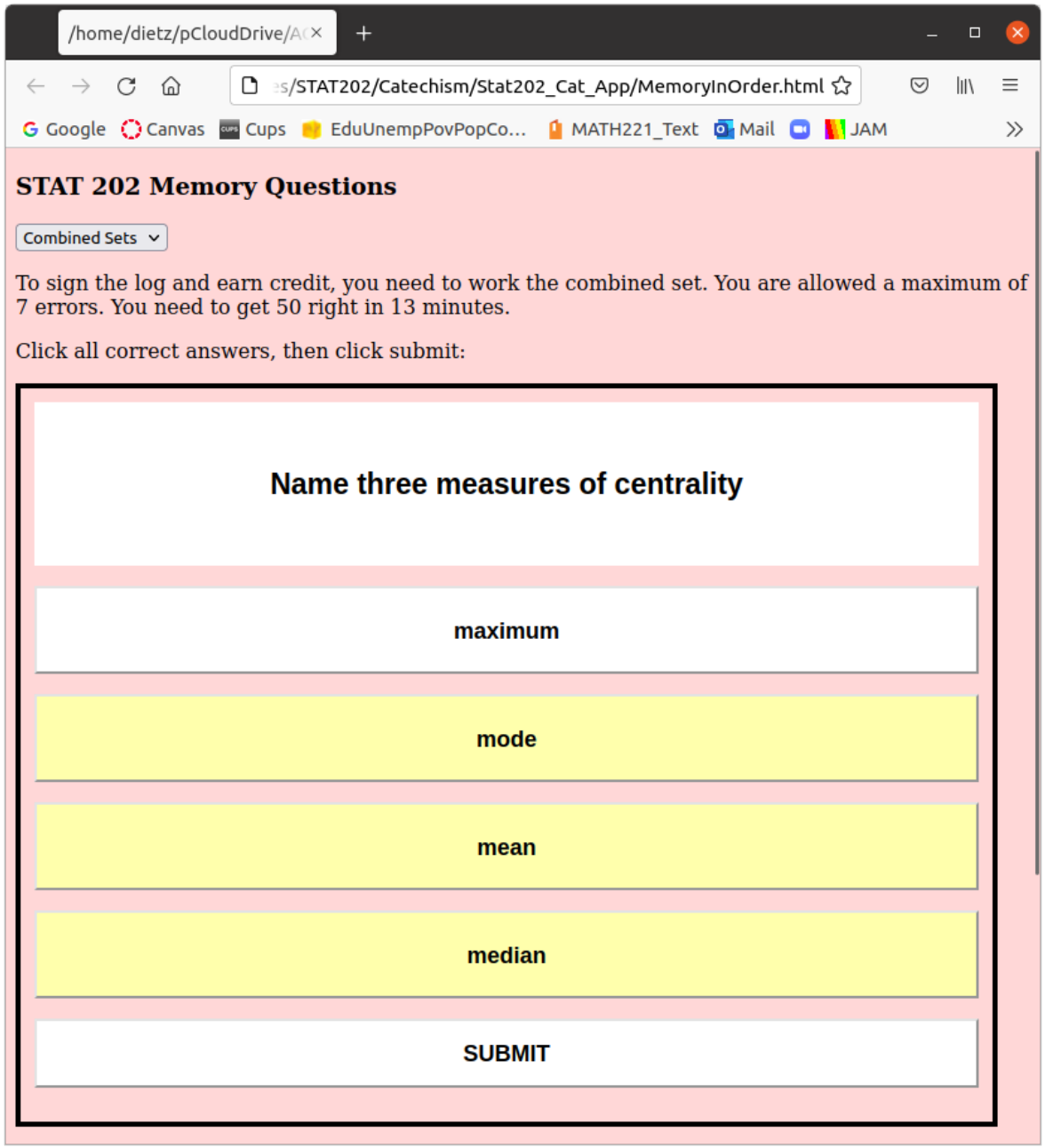
We will soon study this tool in more detail. At this point, you should know:

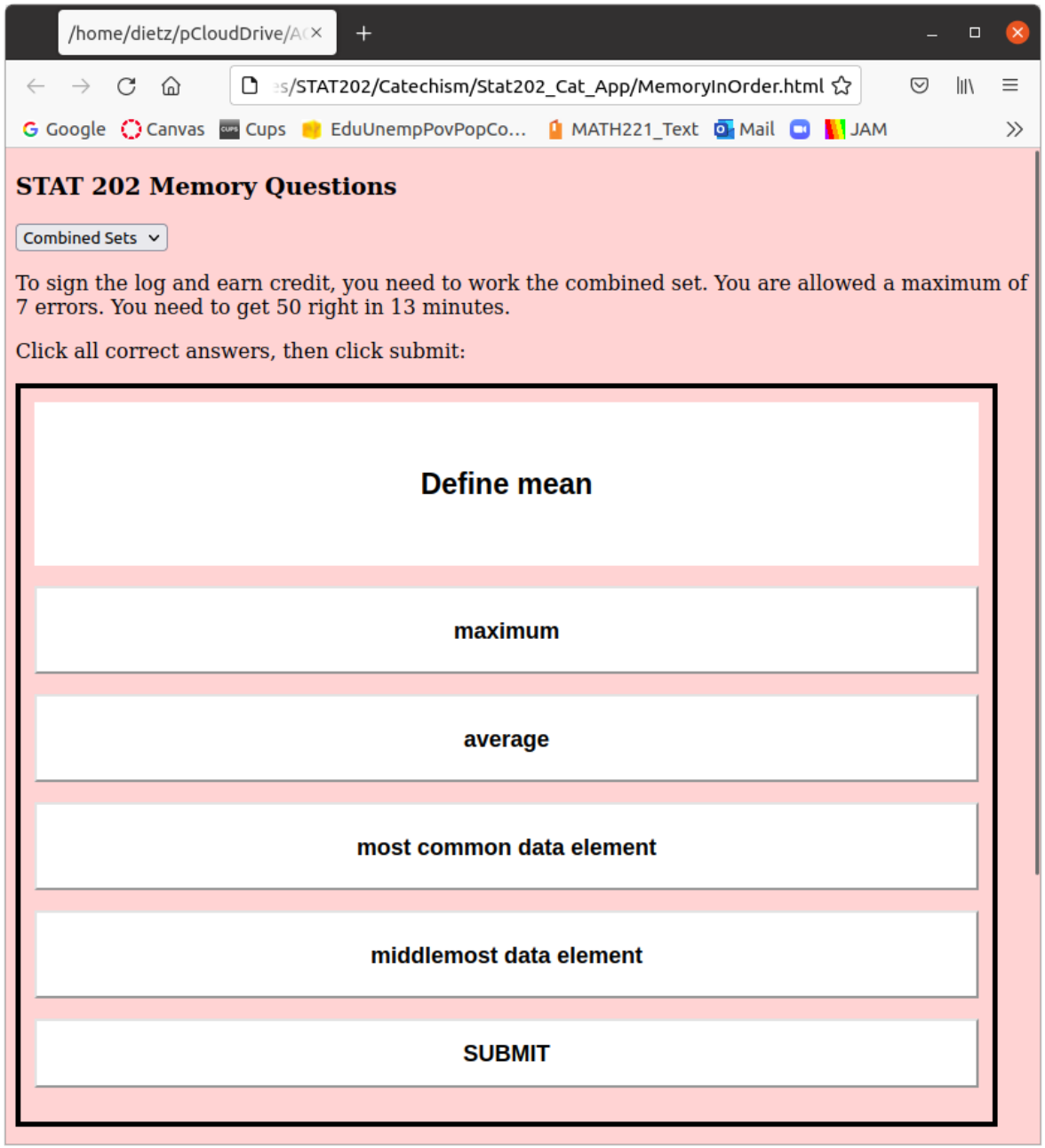
- Like IQR, it indicates the spread of the data set
- Two standard deviations cover about 68% of the data if centered
- $s$ is the *sample standard deviation*
- $\sigma$ (Greek lowercase s) is the *population standard deviation*
- $s$ on StatCrunch is called **Std. dev.**
- $\sigma$ on StatCrunch is called **Unadj. std. dev.**
- For large data sets, $s$ and $\sigma$ are nearly the same.

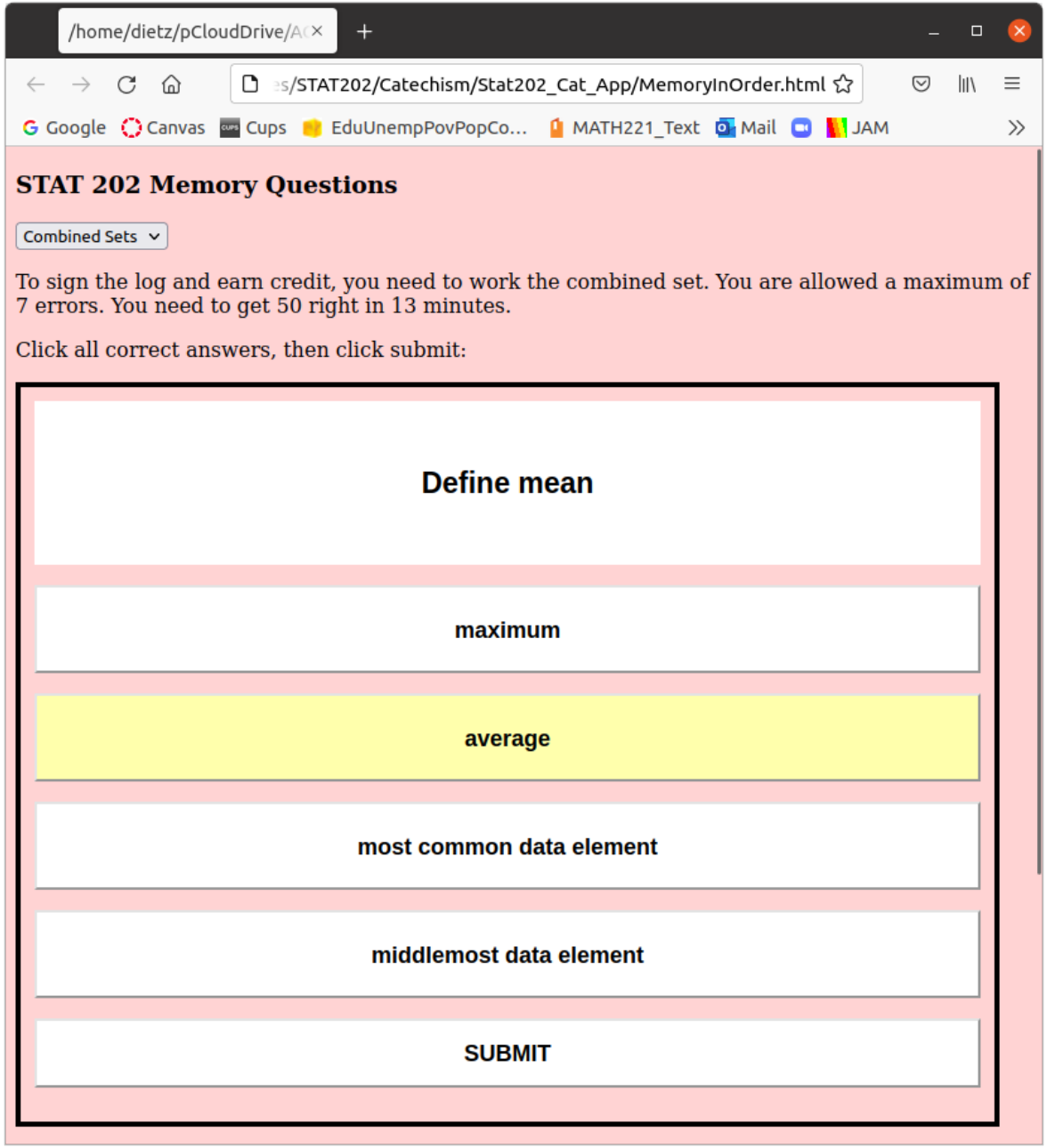
# MEMORY QUESTIONS

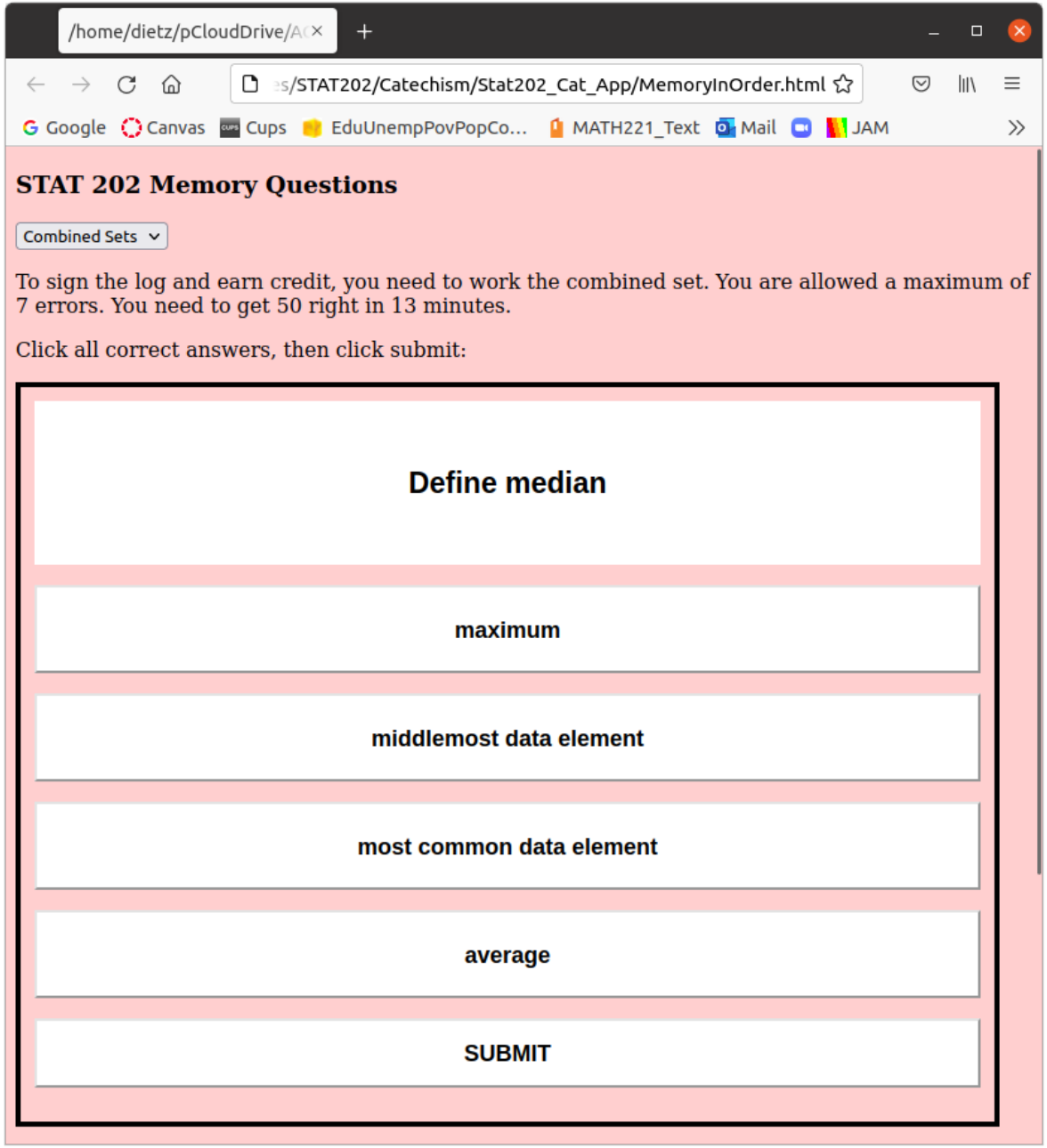
Ten today!

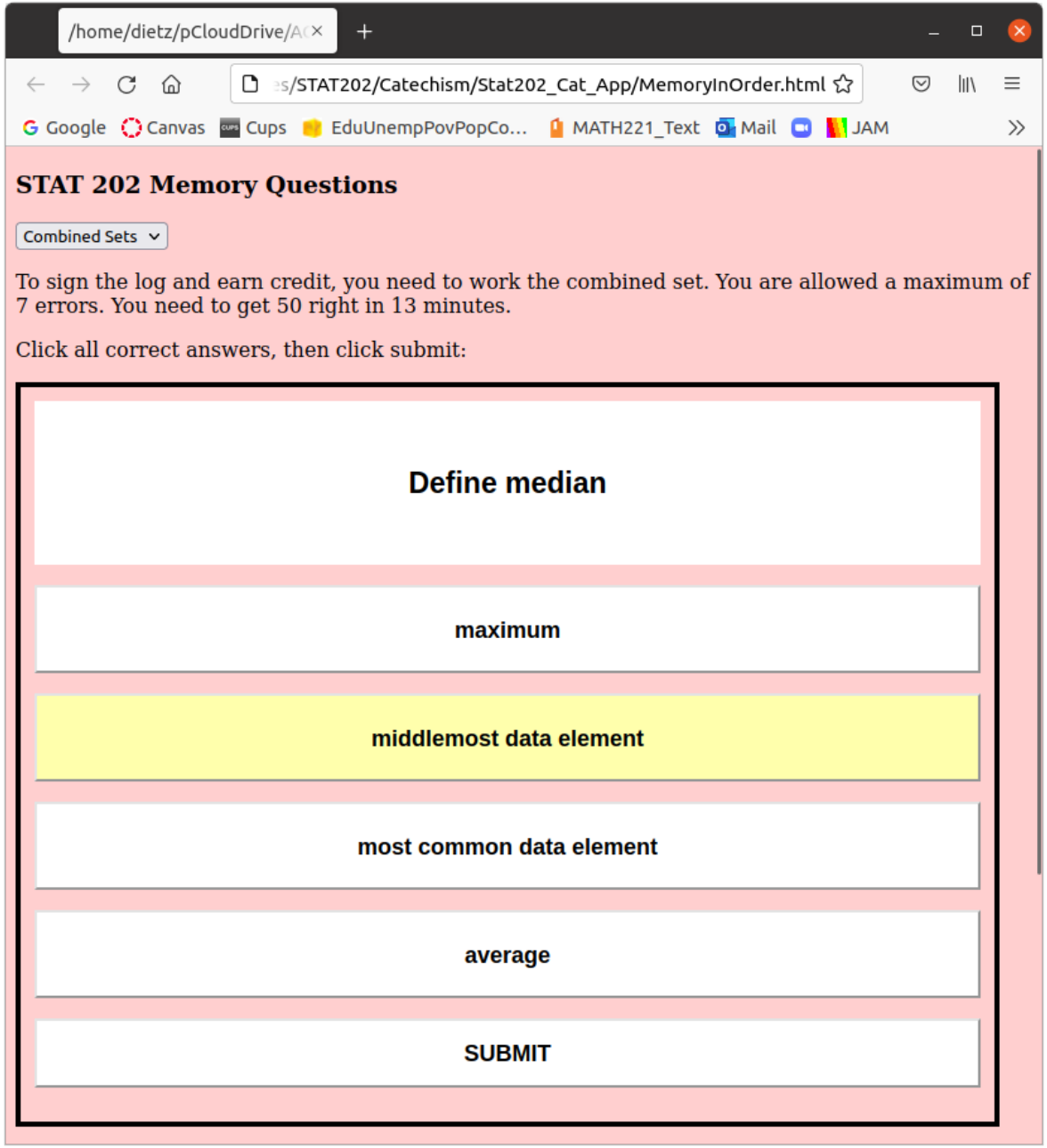

/home/dietz/pCloudDrive/A
/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
Name three measures of centrality
maximum
mode
mean
median
SUBMIT

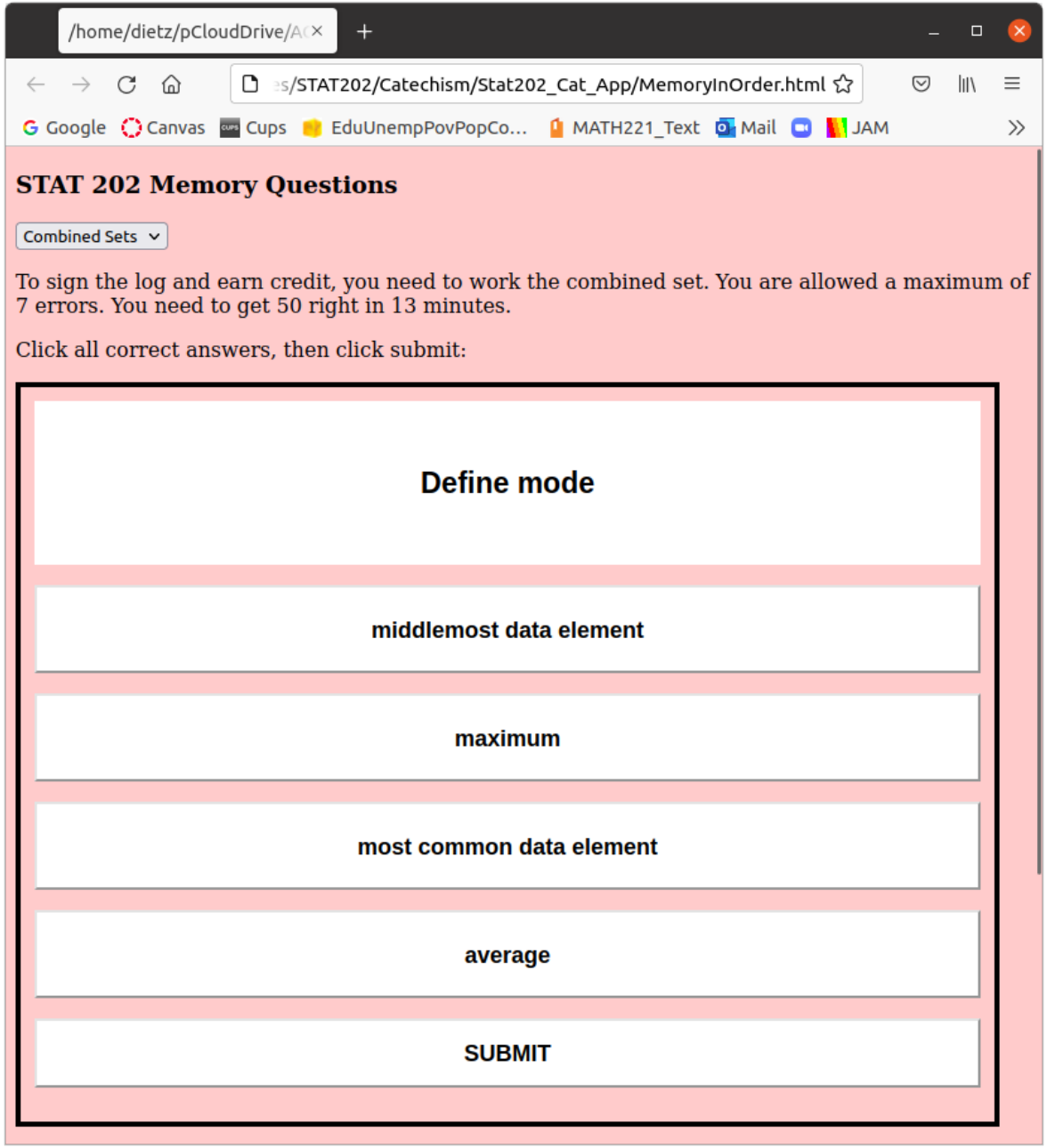

/home/dietz/pCloudDrive/A
/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
Name three measures of centrality
maximum
mode
mean
median
SUBMIT

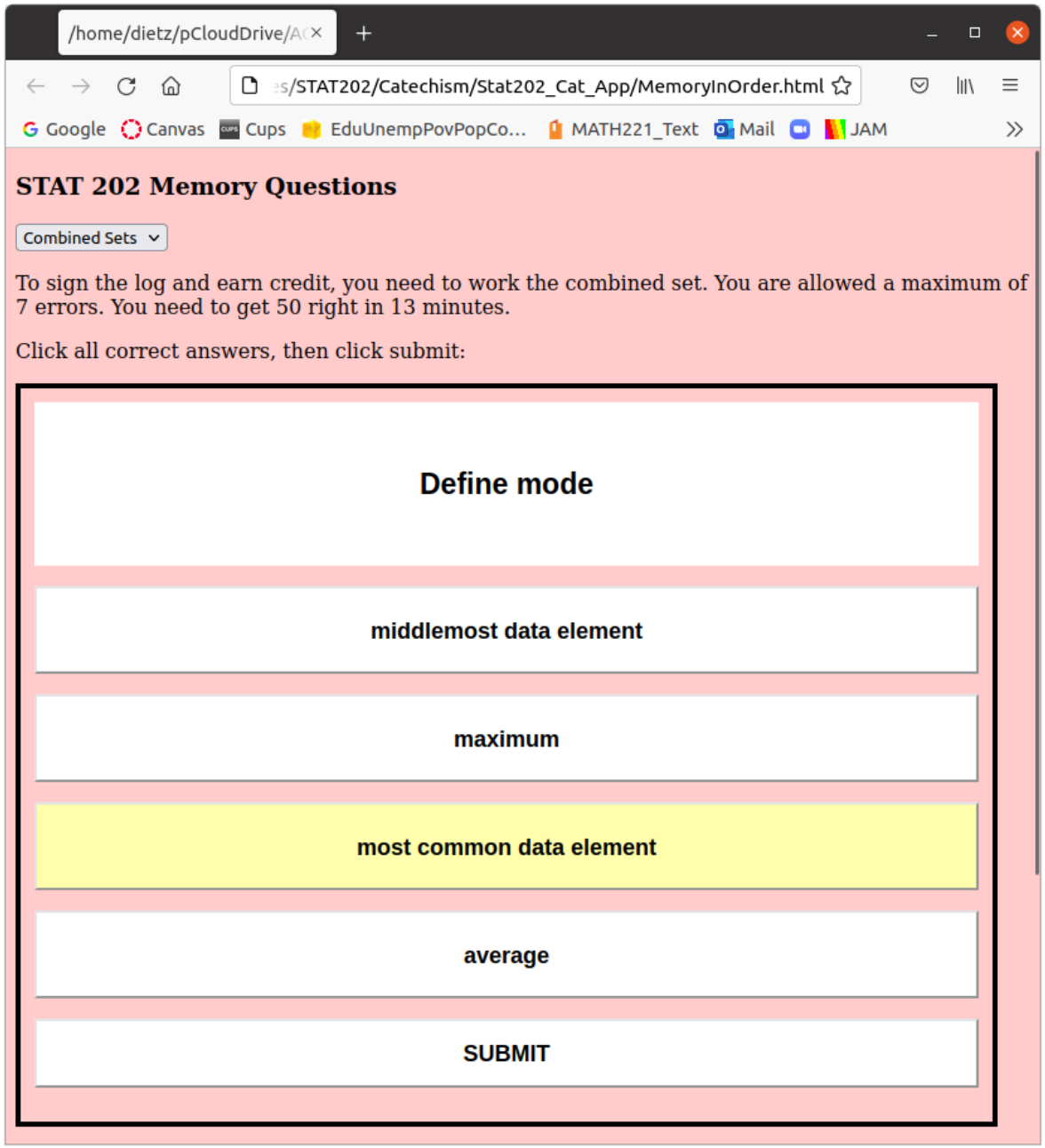

/home/dietz/pCloudDrive/A
/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
Define mean
maximum
average
most common data element
middlemost data element
SUBMIT

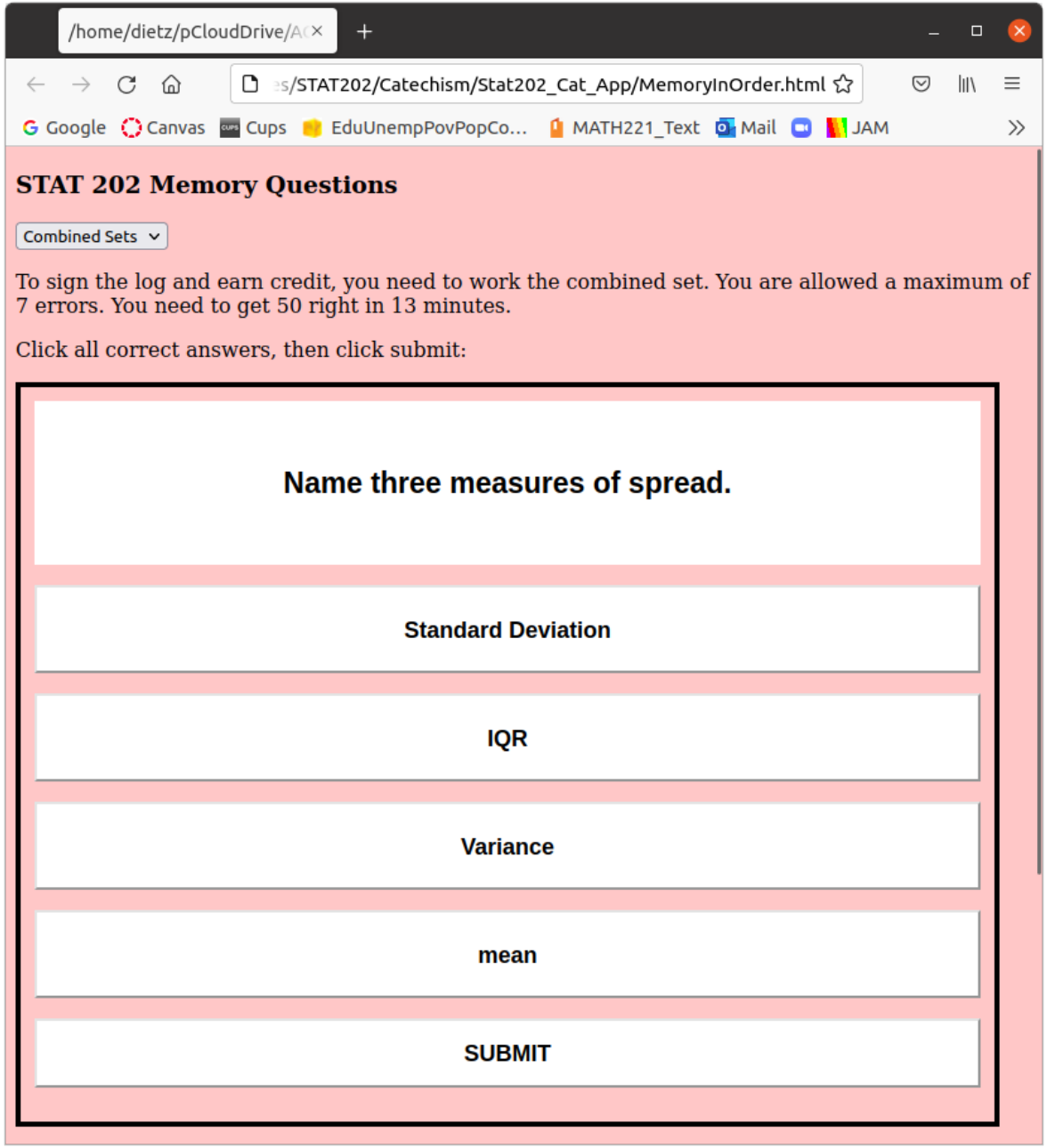

/home/dietz/pCloudDrive/A
/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
Define mean
maximum
average
most common data element
middlemost data element
SUBMIT

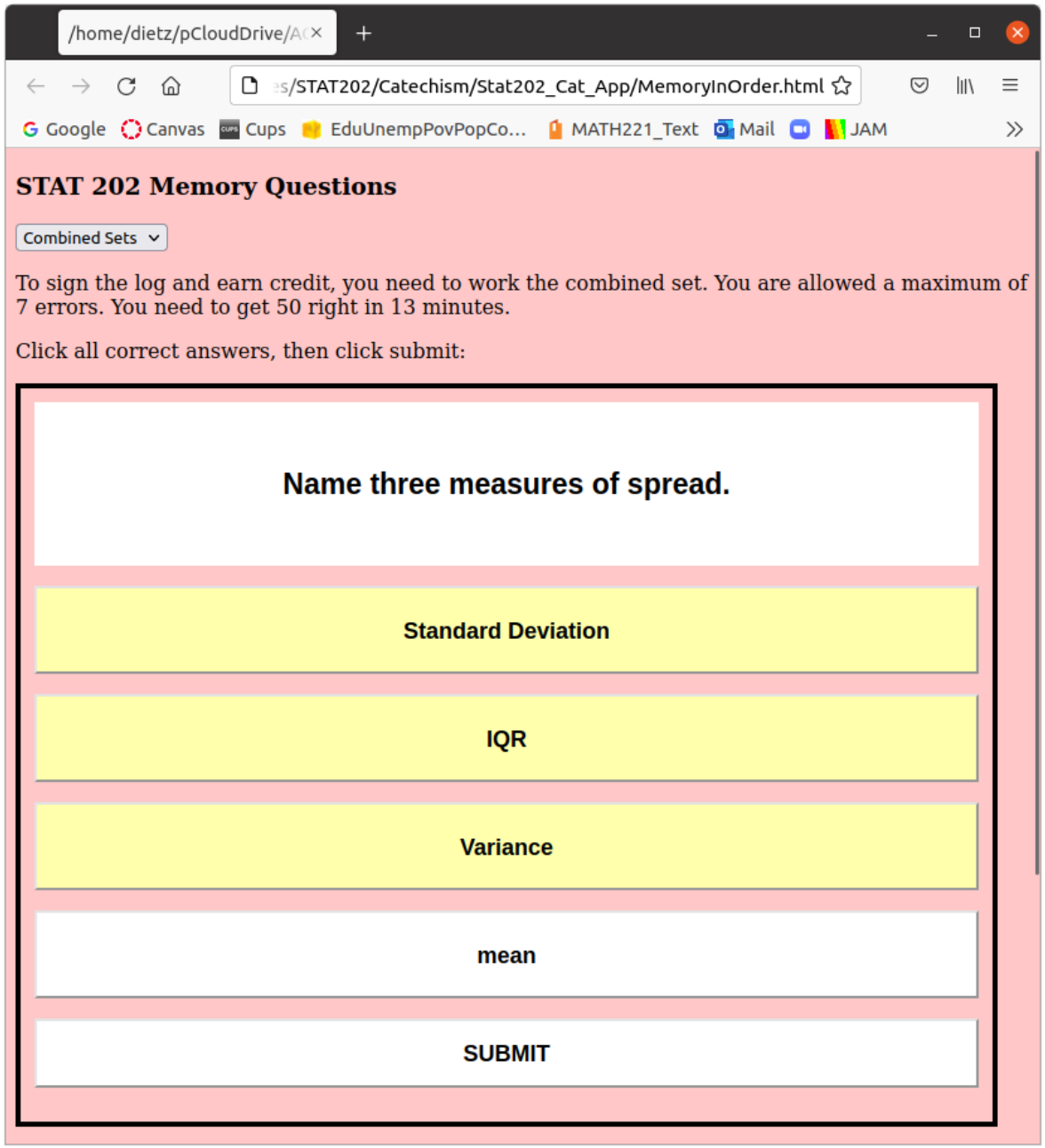

/home/dietz/pCloudDrive/A
/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
Define median
maximum
middlemost data element
most common data element
average
SUBMIT

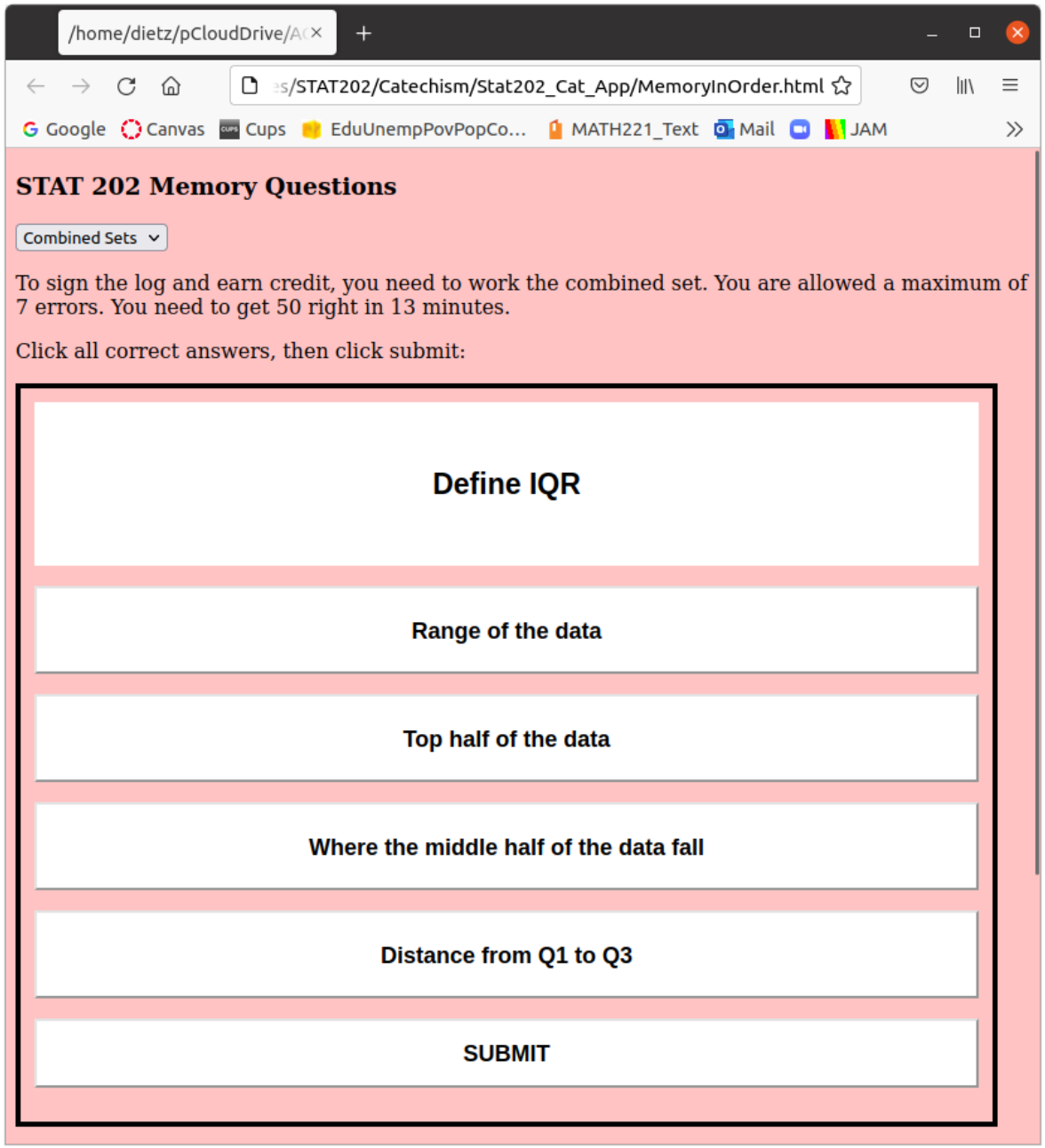

/home/dietz/pCloudDrive/A
s/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
Define median
maximum
middlemost data element
most common data element
average
SUBMIT

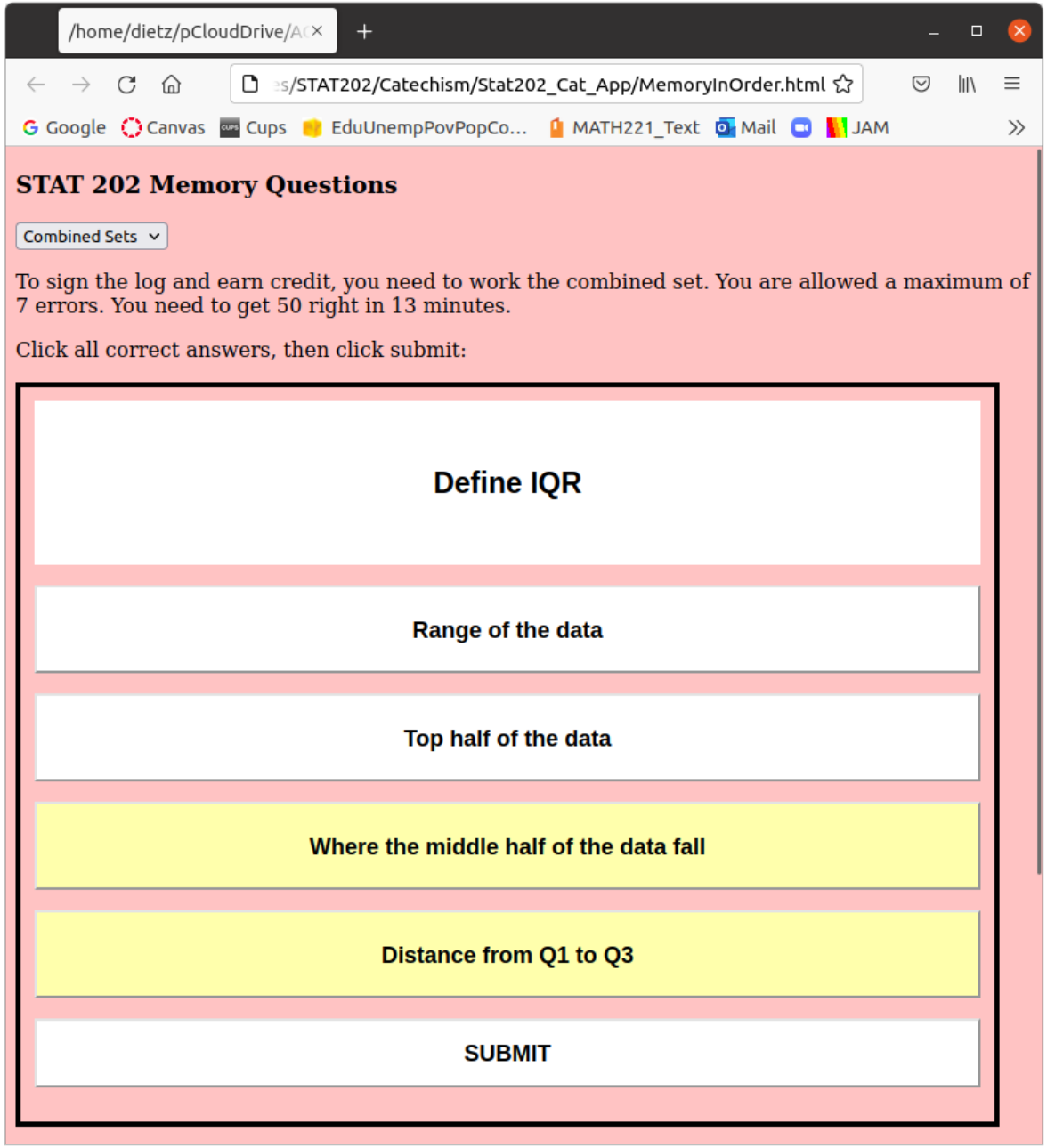

/home/dietz/pCloudDrive/A
/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
Define mode
middlemost data element
maximum
most common data element
average
SUBMIT

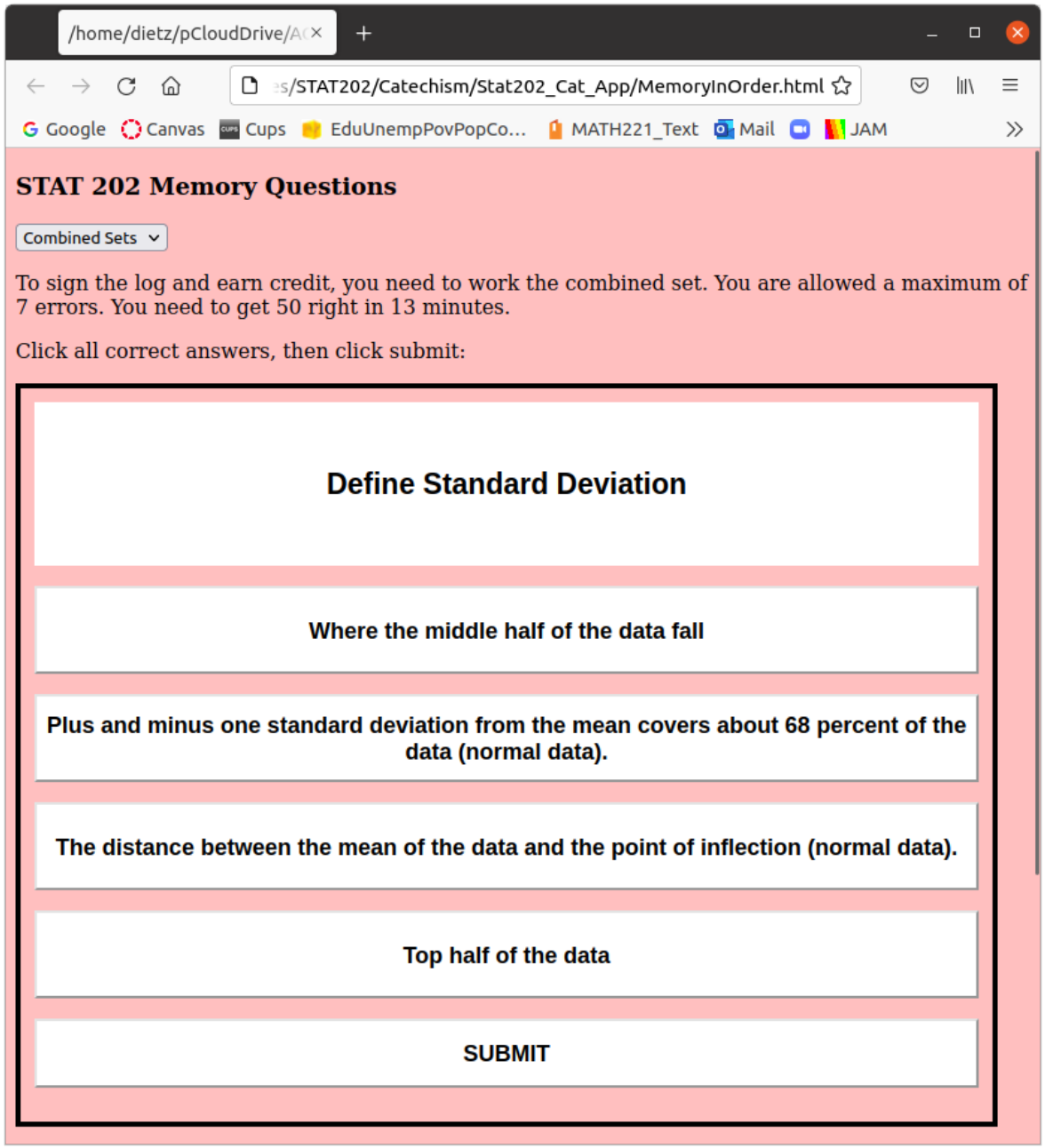

/home/dietz/pCloudDrive/A
/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
Define mode
middlemost data element
maximum
most common data element
average
SUBMIT

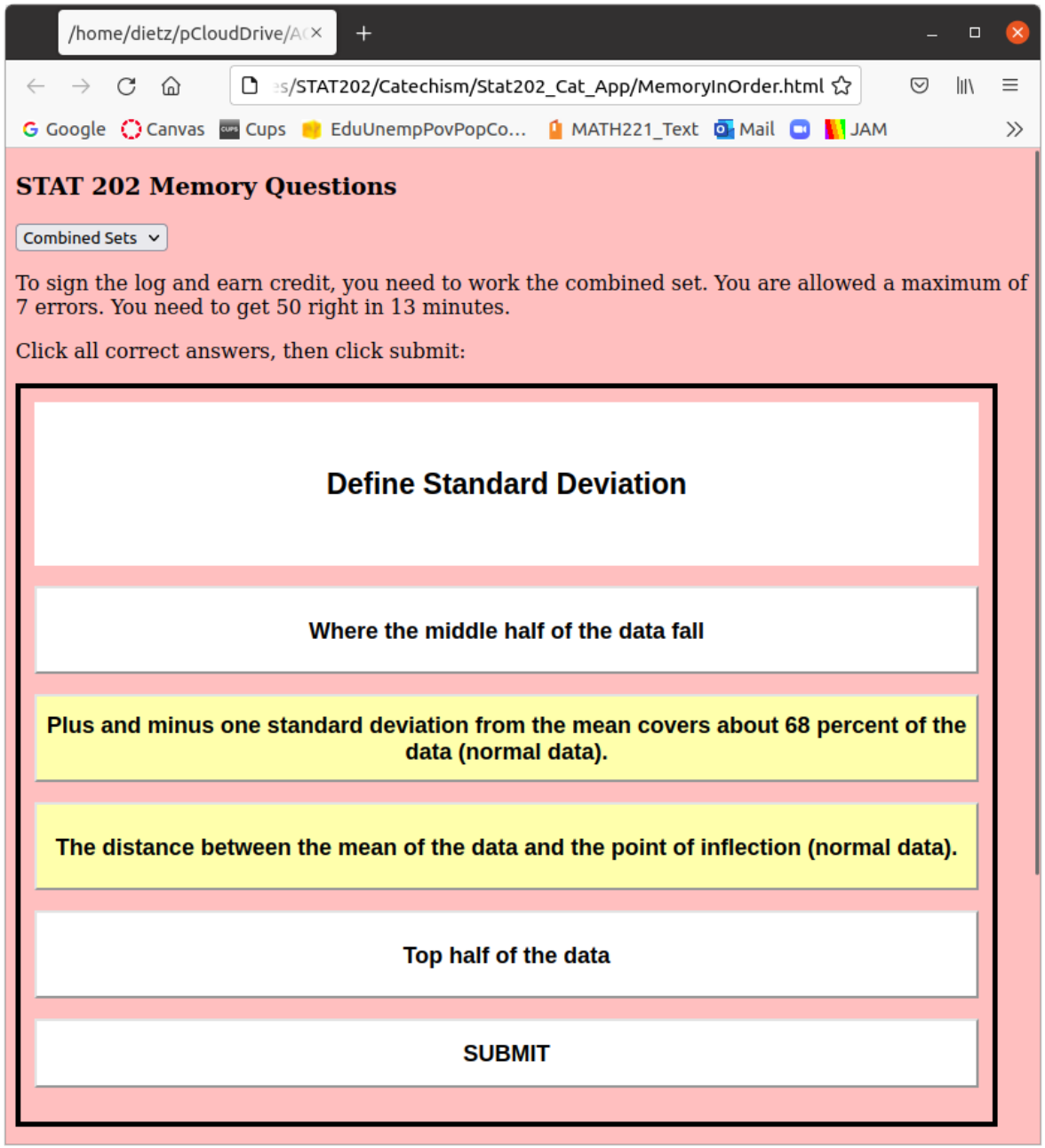

/home/dietz/pCloudDrive/A
/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
Name three measures of spread.
Standard Deviation
IQR
Variance
mean
SUBMIT

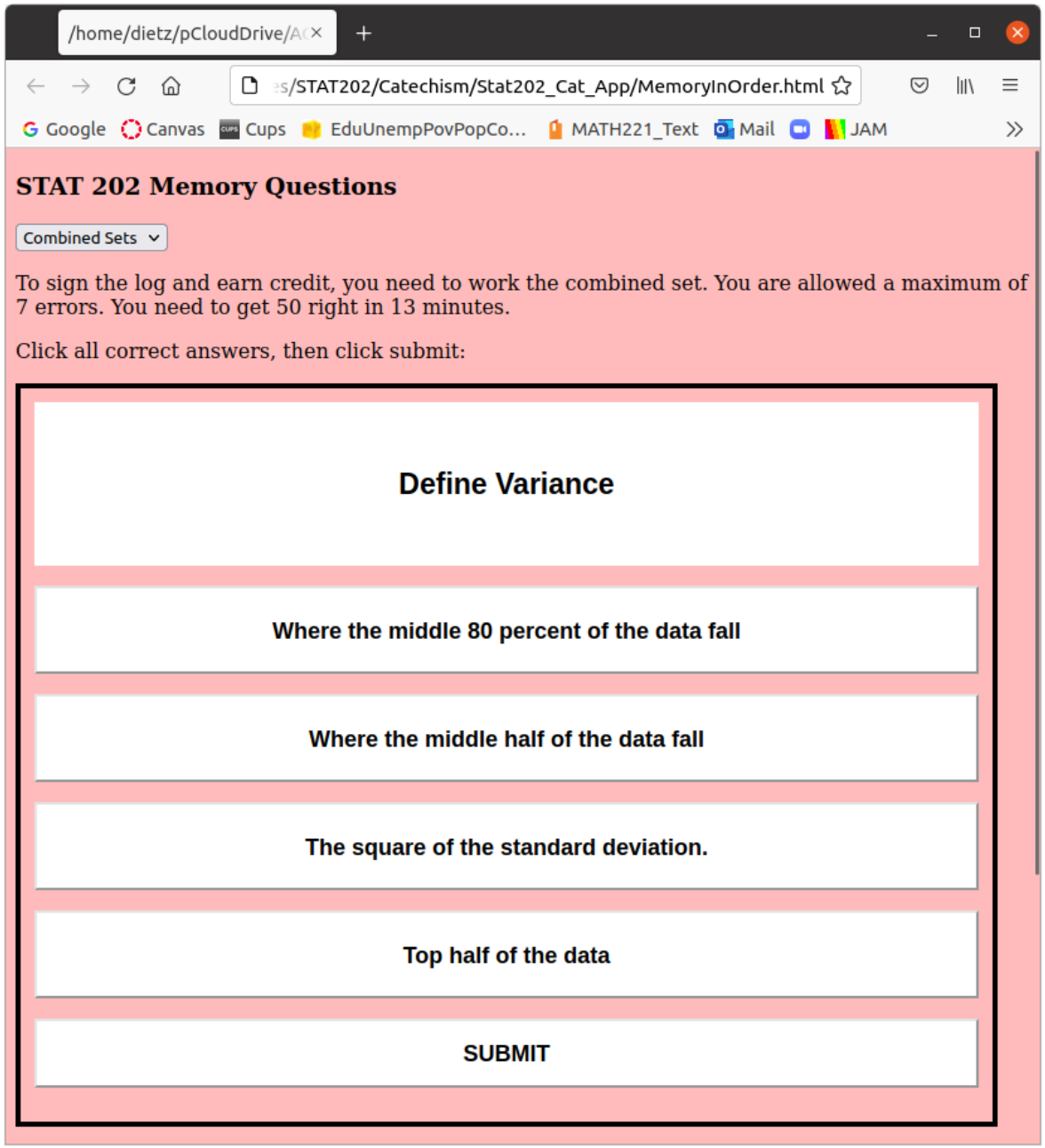

/home/dietz/pCloudDrive/A
/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
Name three measures of spread.
Standard Deviation
IQR
Variance
mean
SUBMIT

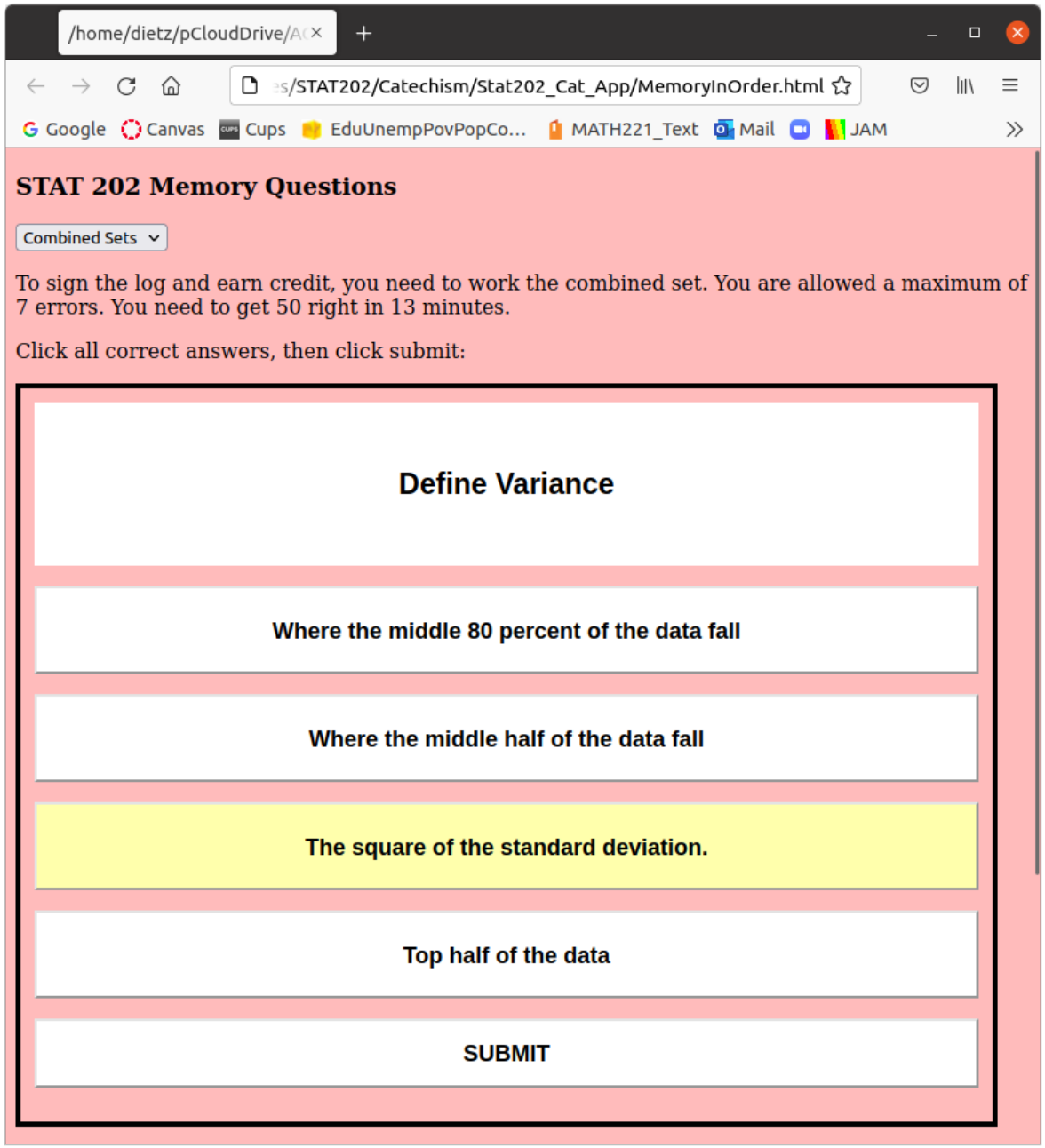

/home/dietz/pCloudDrive/A
/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
Define IQR
Range of the data
Top half of the data
Where the middle half of the data fall
Distance from Q1 to Q3
SUBMIT

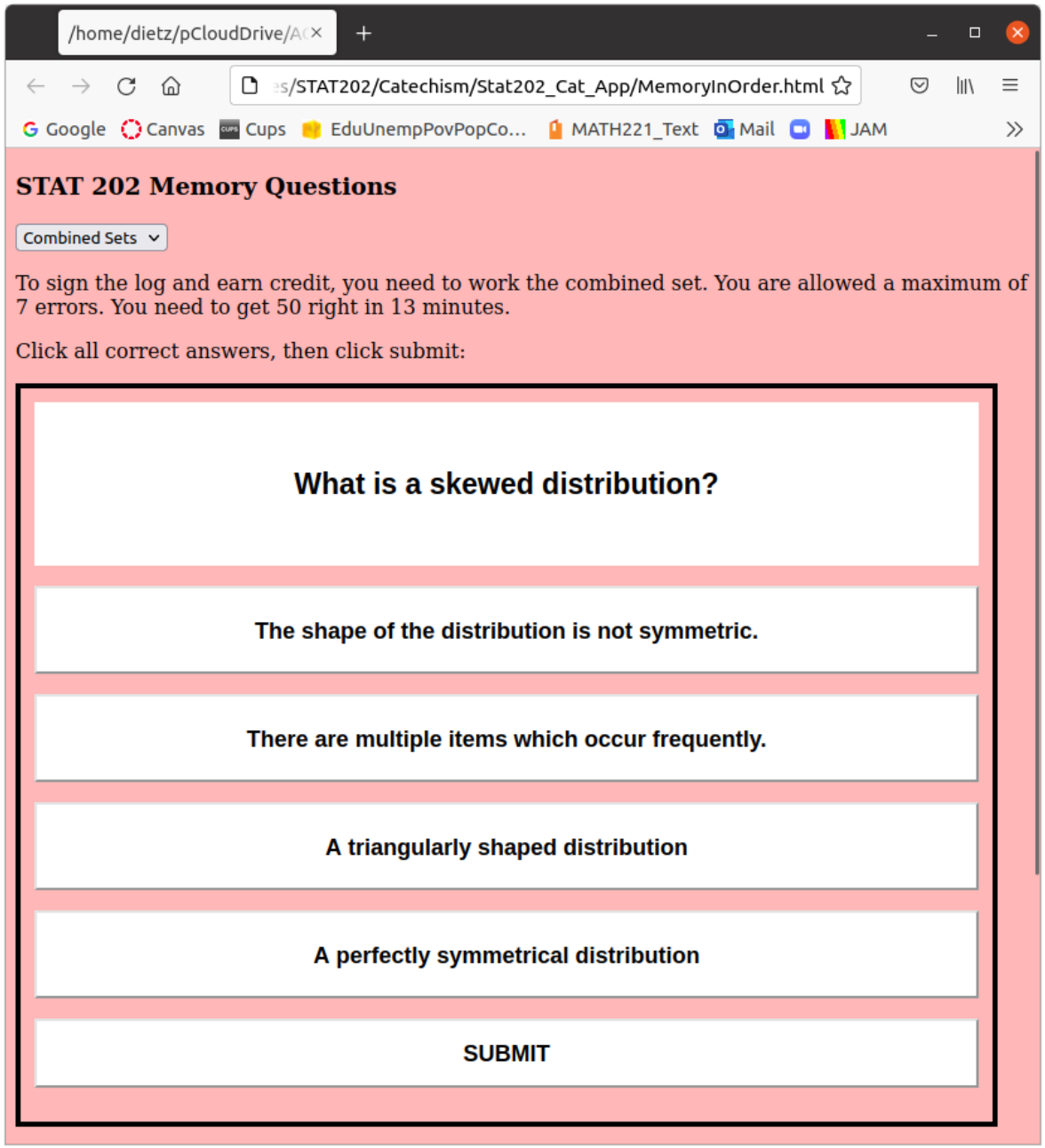

/home/dietz/pCloudDrive/A
/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
Define IQR
Range of the data
Top half of the data
Where the middle half of the data fall
Distance from Q1 to Q3
SUBMIT

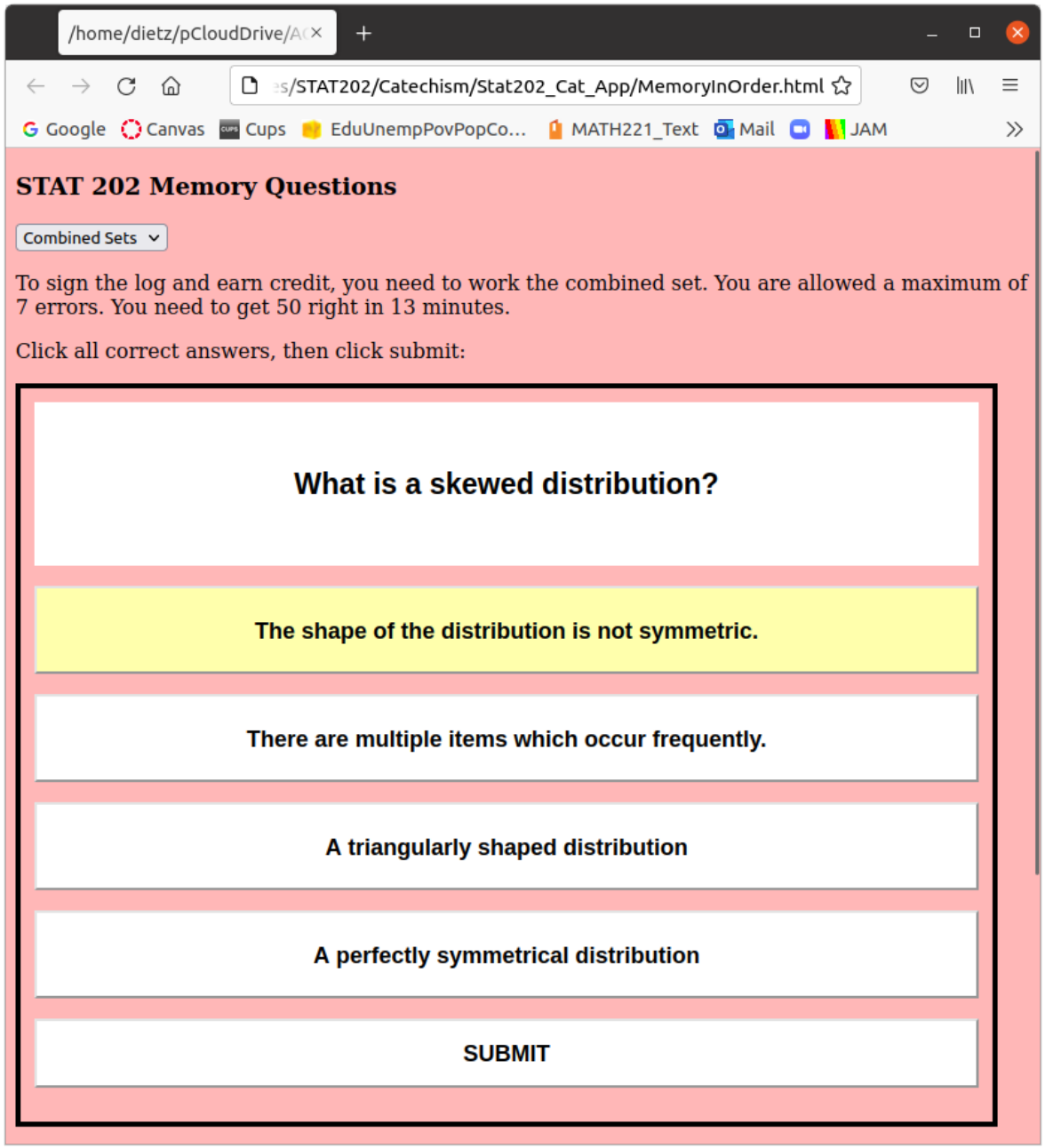

/home/dietz/pCloudDrive/A
/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
Define Standard Deviation
Where the middle half of the data fall
Plus and minus one standard deviation from the mean covers about 68 percent of the data (normal data).
The distance between the mean of the data and the point of inflection (normal data).
Top half of the data
SUBMIT

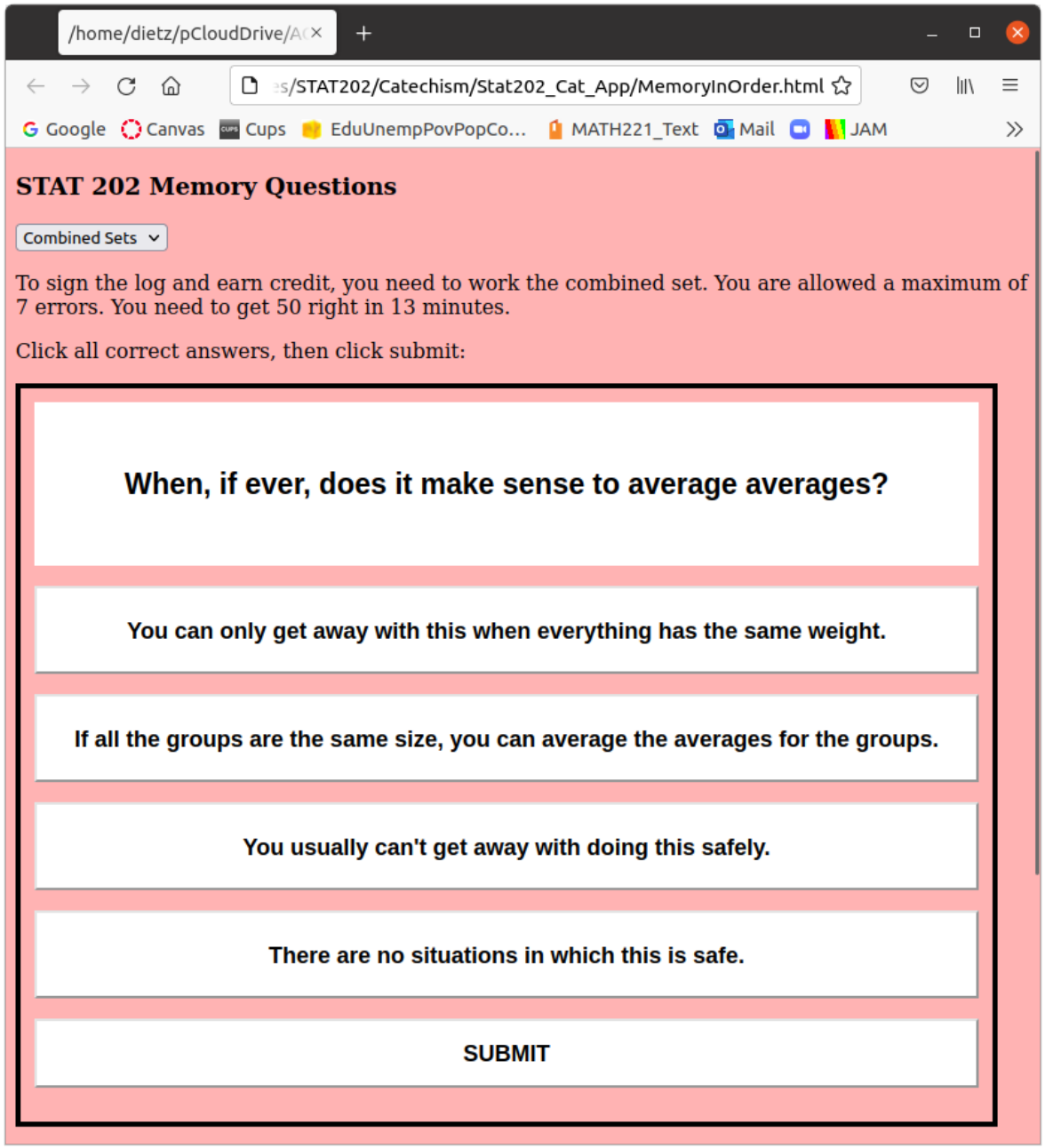

/home/dietz/pCloudDrive/A
/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
Define Standard Deviation
Where the middle half of the data fall
Plus and minus one standard deviation from the mean covers about 68 percent of the data (normal data).
The distance between the mean of the data and the point of inflection (normal data).
Top half of the data
SUBMIT

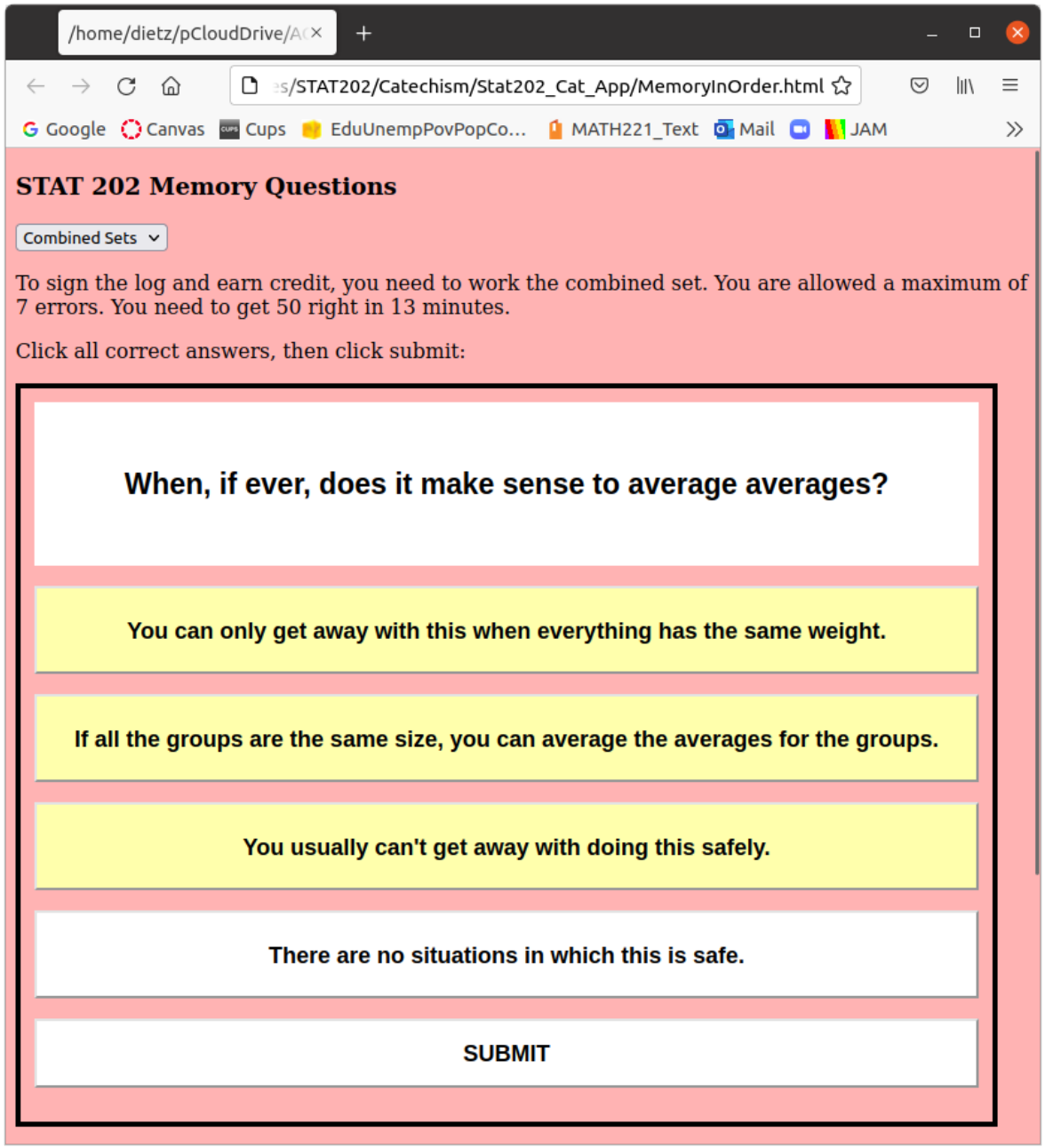

/home/dietz/pCloudDrive/A
/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
Define Variance
Where the middle 80 percent of the data fall
Where the middle half of the data fall
The square of the standard deviation.
Top half of the data
SUBMIT


/home/dietz/pCloudDrive/A
/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:
Define Variance
Where the middle 80 percent of the data fall
Where the middle half of the data fall
The square of the standard deviation.
Top half of the data
SUBMIT


/home/dietz/pCloudDrive/A
/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:
What is a skewed distribution?
The shape of the distribution is not symmetric.
There are multiple items which occur frequently.
A triangularly shaped distribution
A perfectly symmetrical distribution
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:
What is a skewed distribution?
The shape of the distribution is not symmetric.
There are multiple items which occur frequently.
A triangularly shaped distribution
A perfectly symmetrical distribution
SUBMIT


/home/dietz/pCloudDrive/A
/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:
When, if ever, does it make sense to average averages?
You can only get away with this when everything has the same weight.
If all the groups are the same size, you can average the averages for the groups.
You usually can't get away with doing this safely.
There are no situations in which this is safe.
SUBMIT


/home/dietz/pCloudDrive/A
/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:
When, if ever, does it make sense to average averages?
You can only get away with this when everything has the same weight.
If all the groups are the same size, you can average the averages for the groups.
You usually can't get away with doing this safely.
There are no situations in which this is safe.
SUBMIT