Contingency Tables

Donna Dietz

American University

dietz@american.edu

STAT 202 - Spring 2020

Contingency Tables

These notes cover several important tools which are used to compare categorical variables against other categorical variables. These tools work well in conjunction with Chi-Square tests, because we are trying to determine if there is or is not a relationship between categorical variables.

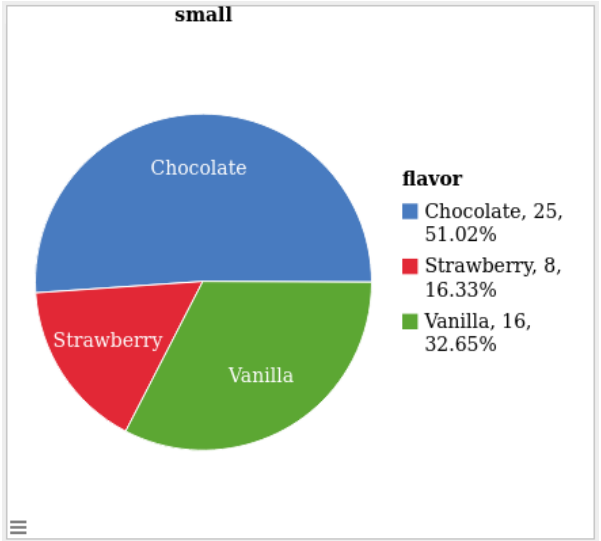
Sets of Pie Charts

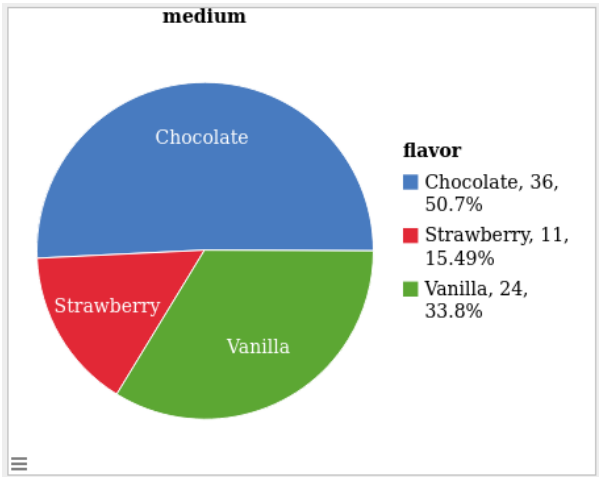
Often, it's useful to look at sets of pie charts to get a sense about your data. Do your pie charts look roughly the same? If they do, there is a good chance there is no statistical significance to the differences in proportions for the different variables.

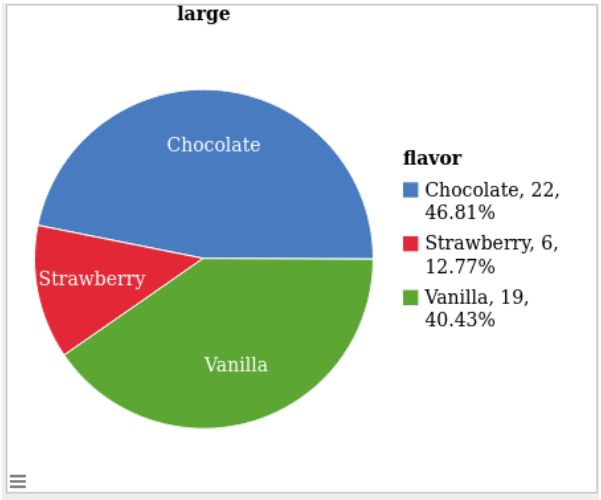
Let's pretend we are selling ice-cream in three flavors and in three sizes. We wonder if the people who buy different sizes tend to buy the flavors in the same proportions as each other. Or, we could state this another way. We wonder if people who like the various flavors tend to buy the sizes in the same proportions regardless of the flavor.

Sets of Pie Charts

flavor	small	medium	large
Vanilla	16	24	19
Chocolate	25	36	22
Strawberry	8	11	6







What do you think?

Sets of Pie Charts

For the ice-cream example, the differences are **not** significant. For now, you can only guess at whether it's significant, but for small samples, you can sanity-check your guesses by asking:

What if just a couple of people changed their decisions?

In this case, we can see that just a few people changing their minds would change the ratios a lot. But, given the numbers we are using, these ratios are pretty close!

Filling in Contingency Tables

	Total	Undergraduate	Graduate
Total			5,735
Men		3,094	2,215
Women		5,029	

Total Student Enrolled In Academic Year 2017-2018 at American University

Fill in the missing values.

(Answer)

	Total	Undergraduate	Graduate
Total	13,858	8,123	5,735
Men	5,309	3,094	2,215
Women	8,549	5,029	3,520

Total Student Enrolled In Academic Year 2017-2018 at American University

Finding Expected Values from Tables

Another tool, besides having side-by-side pie charts (or histograms or bar graphs) is to calculate *expected values* for the cells and compare that against the actual values.

Finding Expected Values from Tables

Independent probabilities multiply

If two variables do not interact (mathematically), we say they are *independent*. If they are *independent*, we can multiply their probabilities together to get the overall probability that they occur together. (And, vice versa!)

For the AU table, let's presume there is no connection between gender and grad/undergrad status.

Expected Counts

Expected Count: Probability times n

The expected count for a pair of traits in a sample (or population) is the number of items in the sample (or population) times the probability that the traits occur together.

Expected Counts

	Total	**Undergraduate**	**Graduate**
Total	13,858	8,123	5,735
Men	5,309		
Women	8,549		

Table: AU enrollment 2017-2018

Example: Undergrads are $\frac{8123}{13858} = 0.5862$ and women are $\frac{8549}{13858} = 0.6169$. So we multiply $0.5862 \times 0.6169 = 0.3616$. We expect 36.16% of the 13858 to be undergraduate women. $0.3616 \times 13858 = 5011.1$ We expect about 5011 undergraduate women. The actual count is 5029. This seems similar to me.

Compute the other three!

Answer

Contingency table results:
Rows: gender
Columns: None

Cell format
Count (Expected count)

	UG	G	Total
Male	3094 (3111.92)	2215 (2197.08)	5309
Female	5029 (5011.08)	3520 (3537.92)	8549
Total	8123	5735	13858

We can do this on StatCrunch to check our work!

Stat > *Tables* > *Contingency* > *With Summary*

Select all columns except labels, then indicate the label column. In Display, choose Expected Count.

Walking and Biking

Columbia, South Carolina	Walked	Biked	Total
Men	8,623	435	9,058
Women	5,125	56	5,181
Total	13,748	491	14,239

This chart is from the Census website. Find the expected value of women biking to work. Is this close to the actual count?

Answer

$$\frac{5181}{14239} \times \frac{491}{14239} \times 14239$$

This is about 178.66 which is **not close** to 56, the actual count!

What would you have thought if your calculation had given 52 and the actual count were 50? Would you think those values were pretty close or not? Would you have concluded that the women were more or less likely to choose biking compared to the men?

Answers

For our data, we had less than a third of the expected number of women choosing to ride their bikes, so it seems that women might be more hesitant to ride bikes to work than men are.

Since we only had two categories for travel and two categories for gender, it really is enough to just test one of the four cells in the two-way table!

We can't use our intuition unless our counts are very close or very far away from our expectations. We will need more tools before we can make meaningful statments for anything else!

Conditional Probability

The *Chance* magazine examined the impact of an applicant's ethnicity on the likelihood of admission to the Houston Independent School District's magnet school programs.

Ethnicity	Admission Decision			
	Accepted	Wait-Listed	Turned Away	Total
Black/Hispanic	485	0	32	**517**
Asian	110	49	133	**292**
White	336	251	359	**946**
Total	**931**	**300**	**524**	**1755**

- What percent of all applicants were White?
- What percent of all applicants were accepted?
- What percent of Black/Hispanic applicants were accepted?

Conditional Probability

When we answered the question "What percent of all Black/Hispanic applicants were accepted?" we answered a conditional probability question.

We could have equally asked, "Given that an applicant was Black/Hispanic, what is the probability they were accepted?"

Answer

$$\frac{485}{517} = 0.9381$$

93.8% of Black/Hispanic applicants were accepted.

The conditional probability of a Black/Hispanic being accepted was 93.8%.

Conditional Probability

Calculating Conditional Probabilities

$$P(A \text{ given } B) = \frac{P(A \text{ and } B)}{P(B)}$$

$$P(A \text{ given } B) = \frac{count(A \text{ and } B)}{count(B)}$$

Conditional probabilities can be found using total counts or probabilities. Disregard all outcomes which have already been eliminated, restrict your attention only to the remaining possible cases, and your results will be correct.

Some definitions

Population versus Sample

A population is an entire group you would like to study. The sample is a selection from that population.

Sample versus Census

If you are lucky enough to get ahold of all the data for your entire population, then your sample **is** your population and you call that a **census**.

More definitions

Parameters

Parameters are numerical values associated with populations. These are generally represented with Greek letters like μ for average and σ for standard deviation. We use p for a proportion in a population. These are values we wish we could magically know!

Statistics

The word *statistics* has many different definitions. When it is being used in this way, in contrast to a parameter, it means the numerical values associated with a sample. Usually we are trying to guess a parameters by looking at the statistics from the sample. We tend to use Roman letters. We use s for the so-called *sample standard deviation* which is **not** the standard deviation **of** the sample. It is the estimator for the population's standard deviation. However for $\bar{x}$, we are lucky. The estimator for the popuation mean happens to be calculated the same way as μ.

Bias

Since $\bar{x}$ and μ are both calculated by adding up all the data you have and dividing by the amount of data you have, you won't see two different buttons on a calculator or two different things to calculate on StatCrunch. This is because the *mean* is **not** a biased estimator!

However, there is a very easily understood biased estimator. The maximum of a sample is rarely as high as the maximum of a population. If you asked the students in a class to write down their SAT scores on slips of paper and share them anonymously with the rest of the class, it's unlikely that 1600 would appear. It might, but it's not highly likely. However, we all know that some people in the population manage to get that coveted perfect score. So, this is a concrete example of a biased estimator.

Bias

The other kind of bias:

In the context of a survey, bias means what you probably think it means: A biased survey is often written to encourage a certain response. Other issues include "who" is doing the survey, where it is done and so forth.

Why s and σ

If we were to use the formula for σ to estimate s, we would consistently be a bit on the low side. This means that sometimes we would be too high and sometimes too low, but overall, we would be too low more often than we would be too high. This is why σ can't be used to estimate the standard deviation of the population. This is why we need to use s. Luckily, the formula for s is very similar to the one for σ, and even luckier, StatCrunch doesn't mind doing either one!

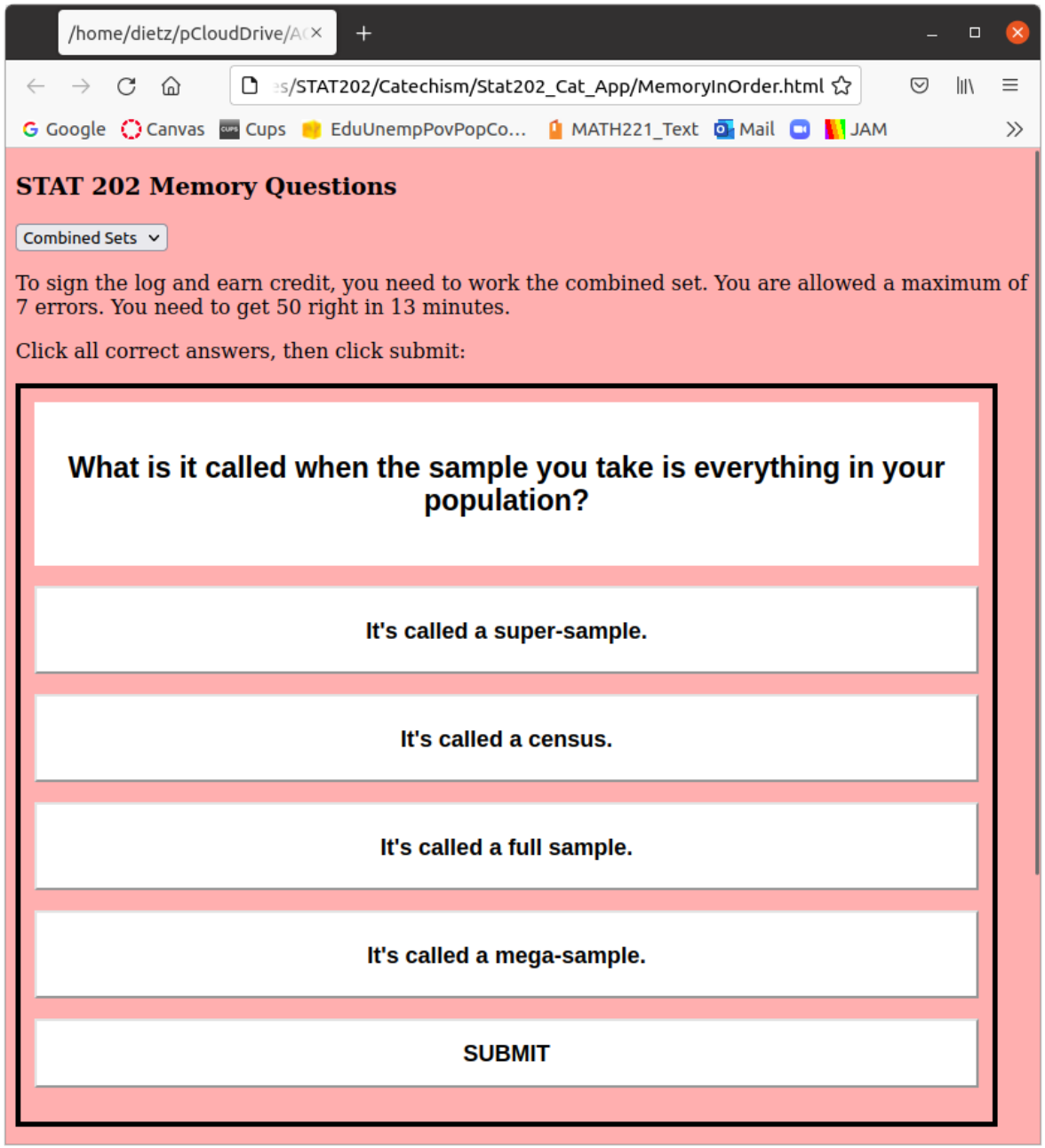
Caution!

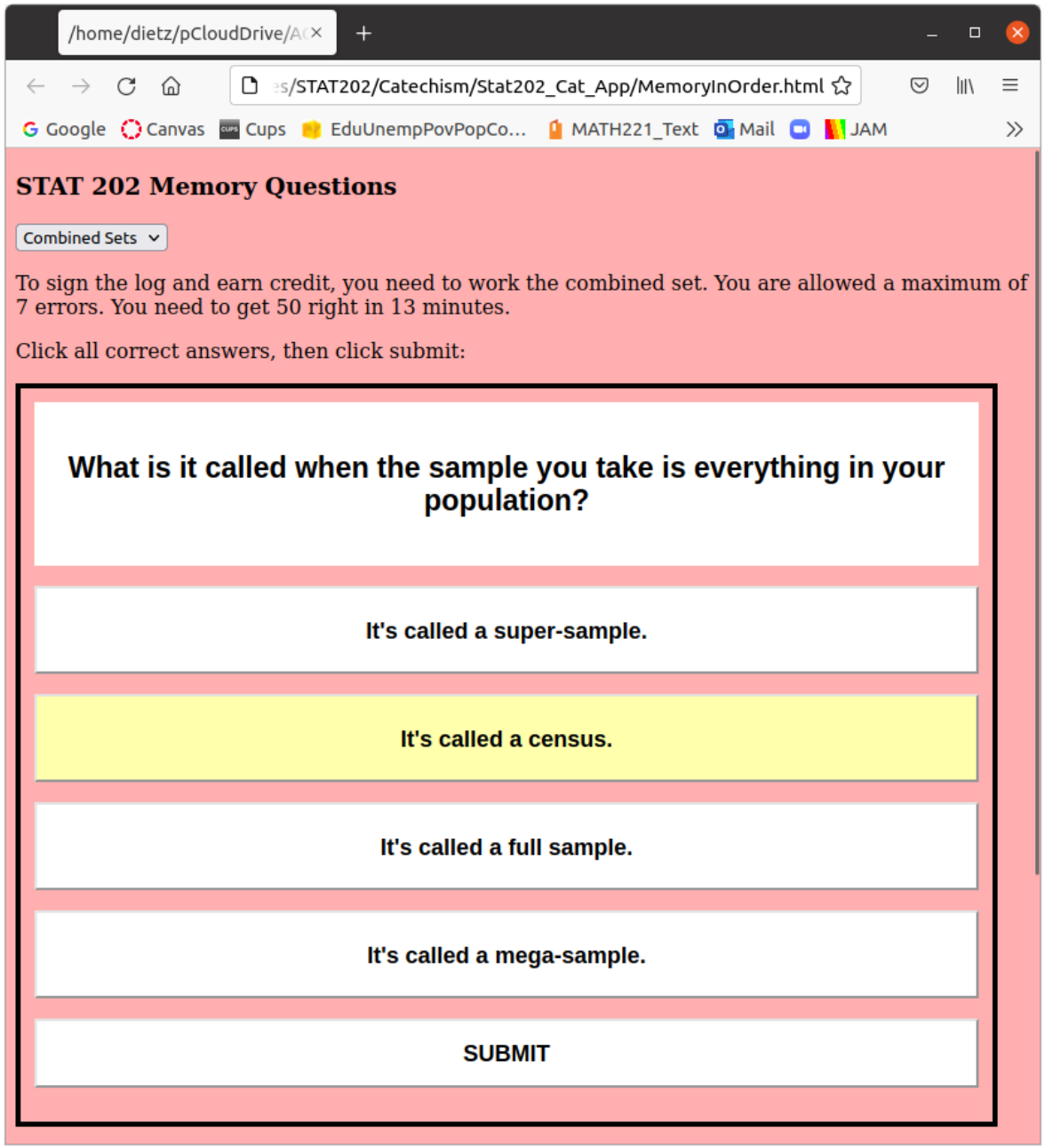
When looking at two-way tables, be very aware of whether the counts (or percents) you are looking at are in reference to an entire population or just a sample. If you are looking at an entire population, you can make more solid statements. You can say things like "54.1% of the student body is female." if you have actual counts. If you only have a sample, you should adjust your language accordingly. For example, "In our study, 52% of respondents said they preferred chocolate."

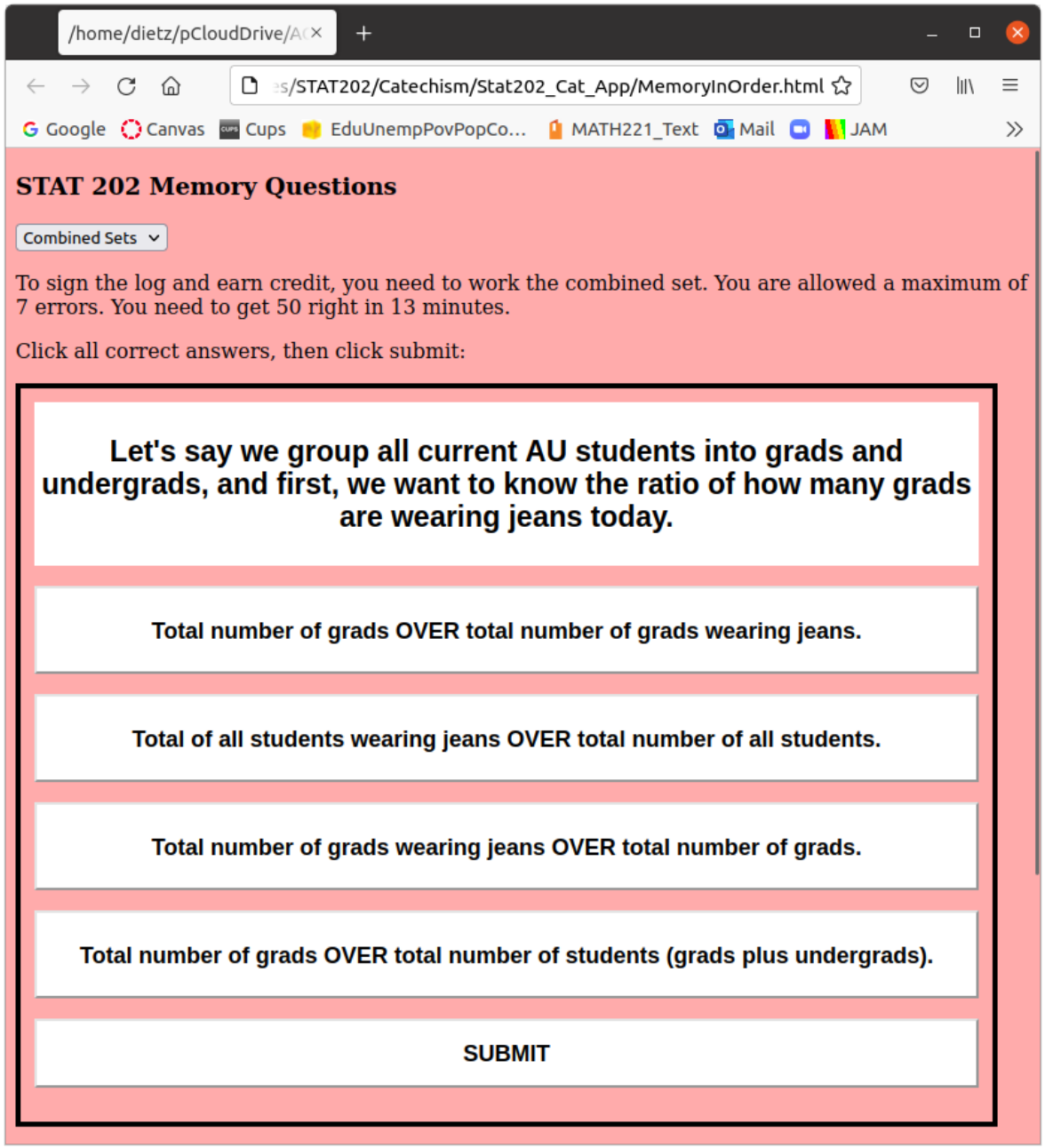
Later in the semester, we will learn how to correctly say more than this.

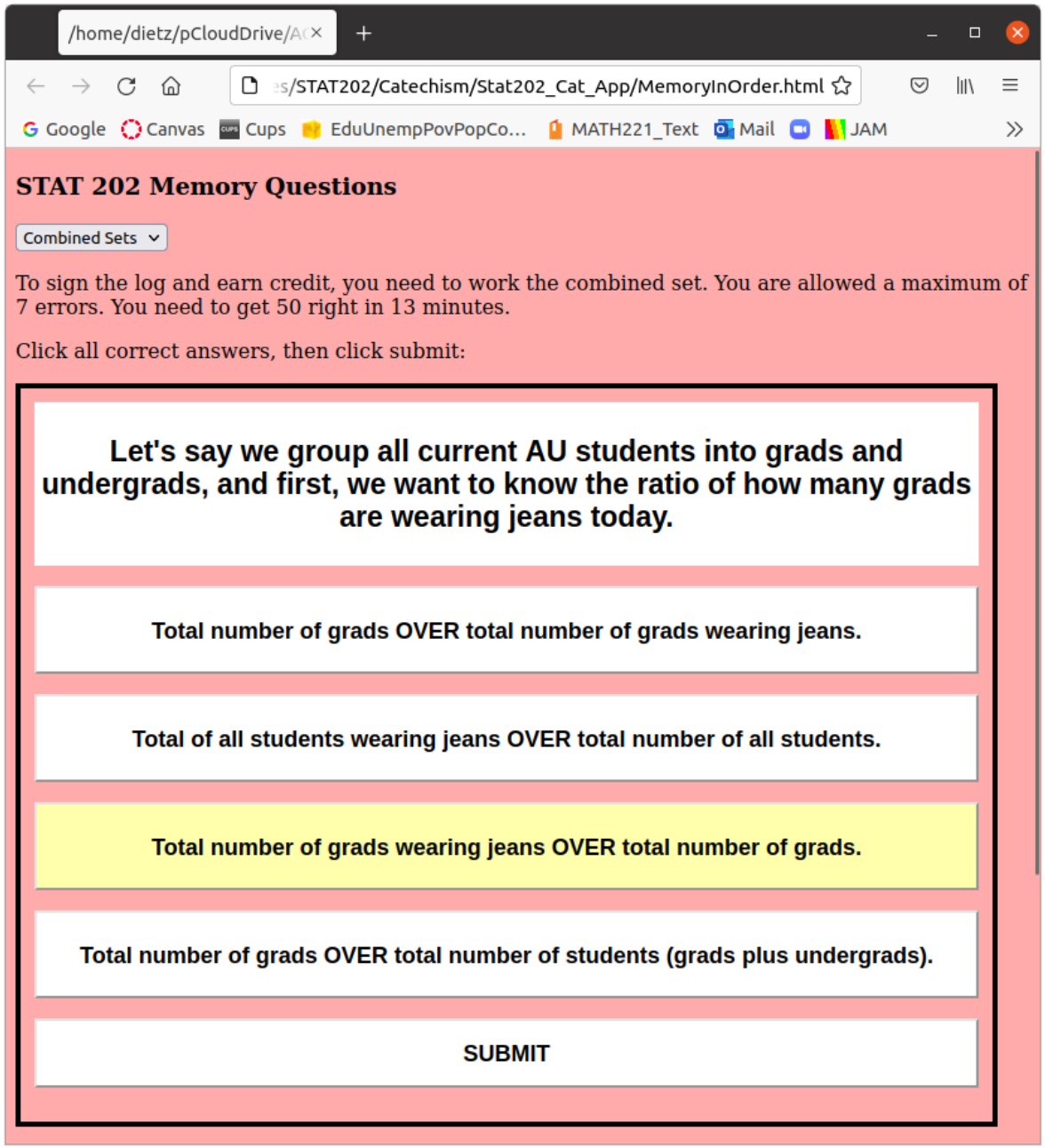
MEMORY QUESTIONS

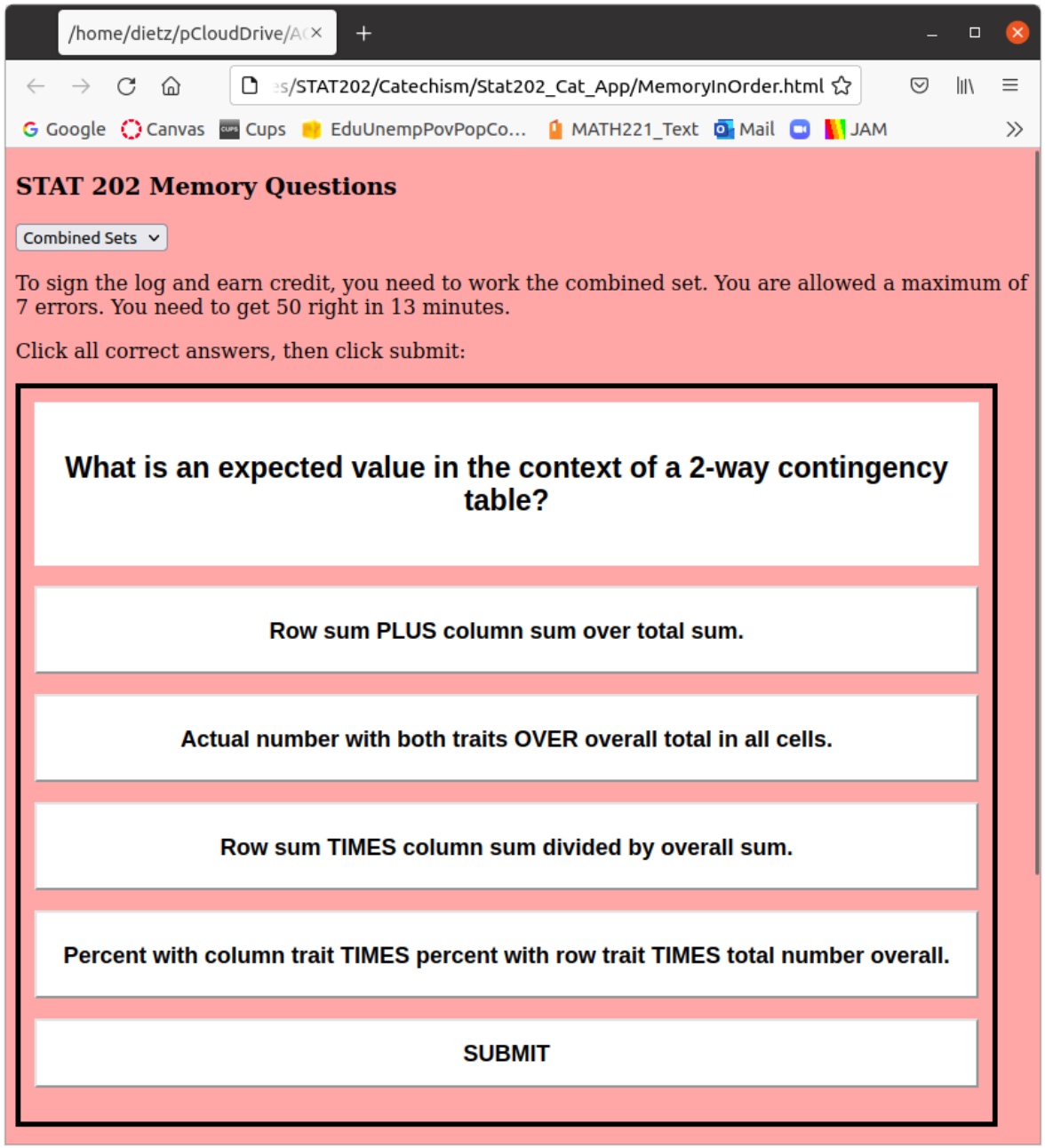
Five today!

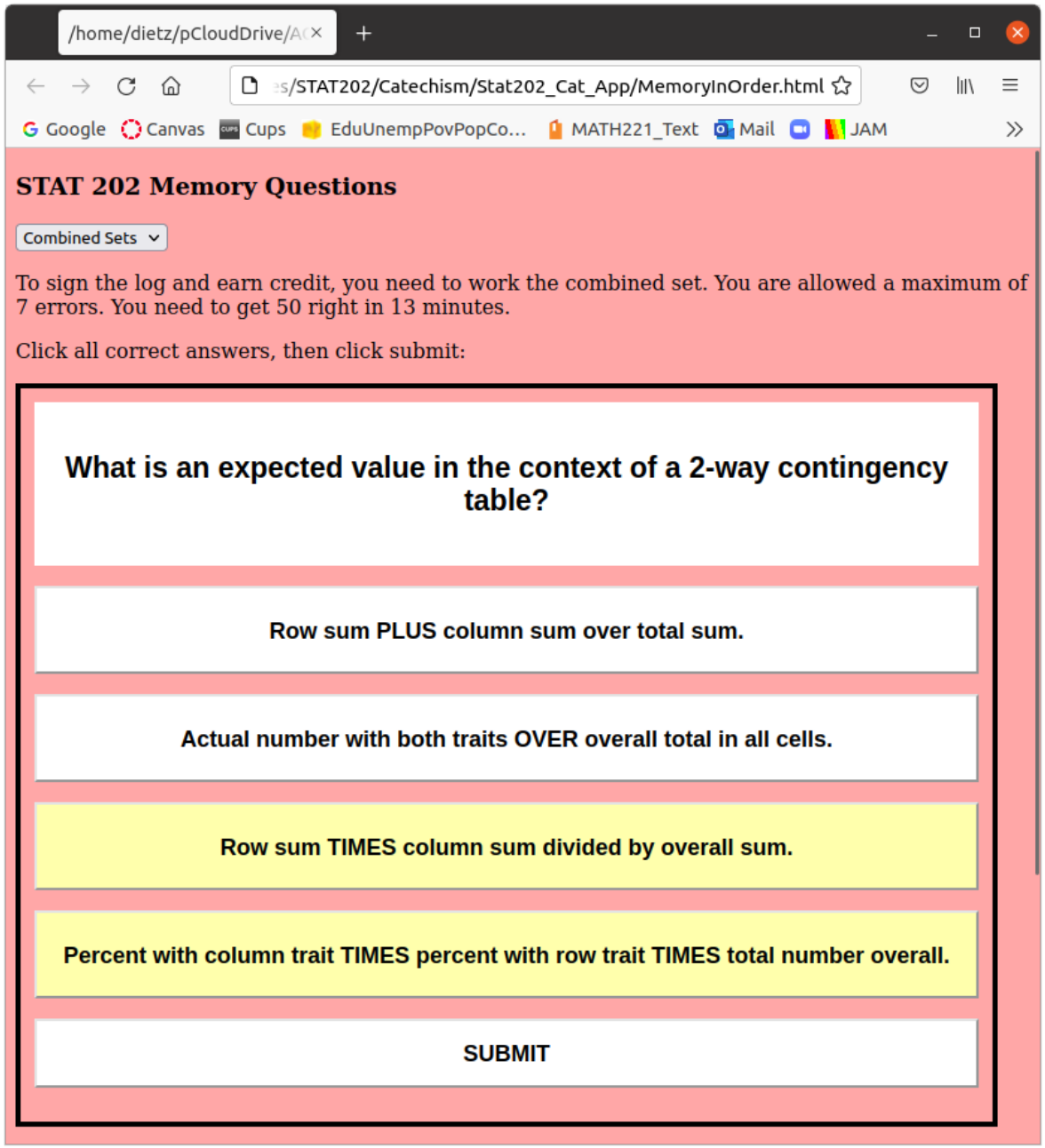

/home/dietz/pCloudDrive/A
/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
What is it called when the sample you take is everything in your population?
It's called a super-sample.
It's called a census.
It's called a full sample.
It's called a mega-sample.
SUBMIT

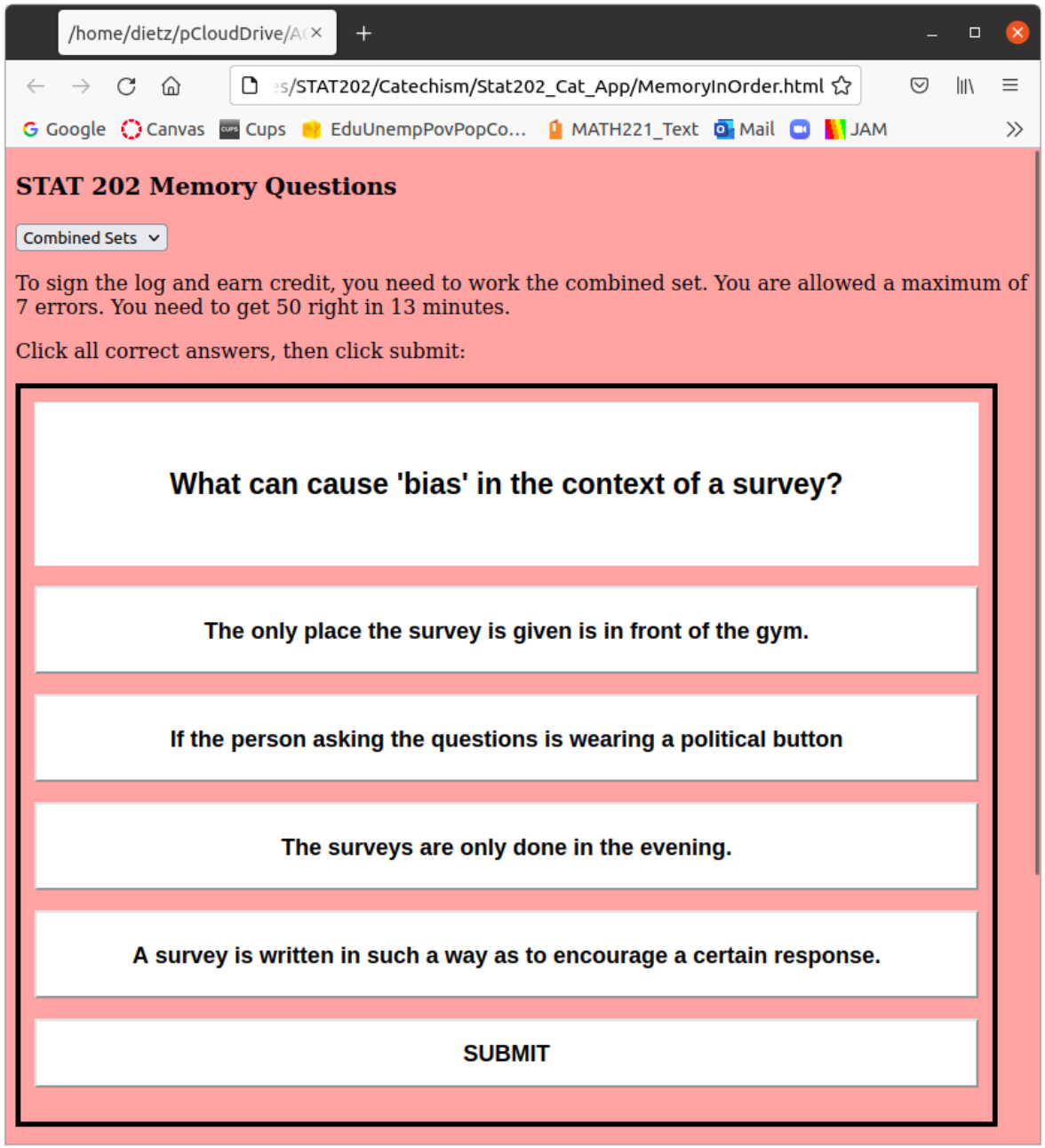

/home/dietz/pCloudDrive/A
/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
What is it called when the sample you take is everything in your population?
It's called a super-sample.
It's called a census.
It's called a full sample.
It's called a mega-sample.
SUBMIT

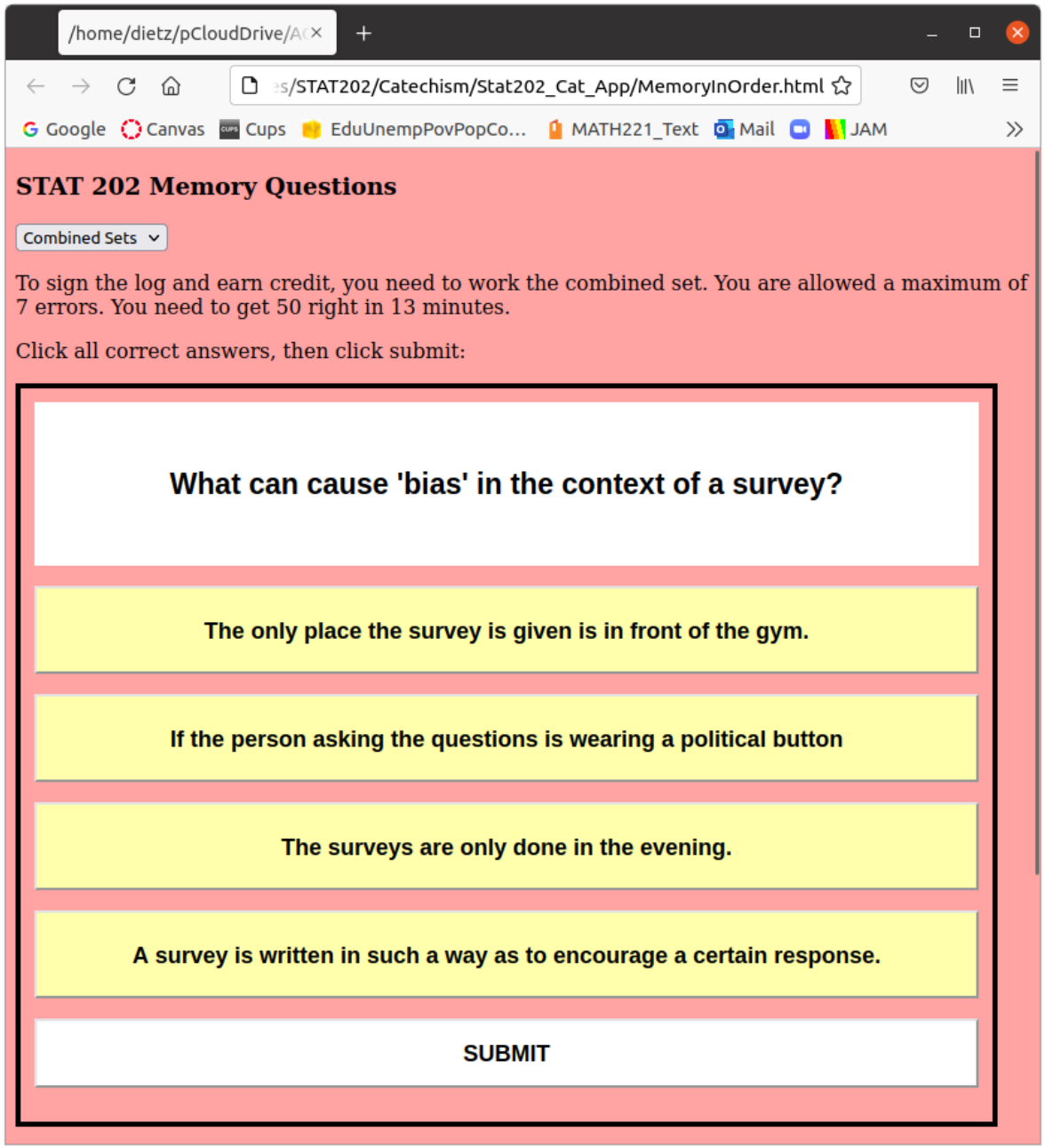

/home/dietz/pCloudDrive/A
s/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
Let's say we group all current AU students into grads and undergrads, and first, we want to know the ratio of how many grads are wearing jeans today.
Total number of grads OVER total number of grads wearing jeans.
Total of all students wearing jeans OVER total number of all students.
Total number of grads wearing jeans OVER total number of grads.
Total number of grads OVER total number of students (grads plus undergrads).
SUBMIT

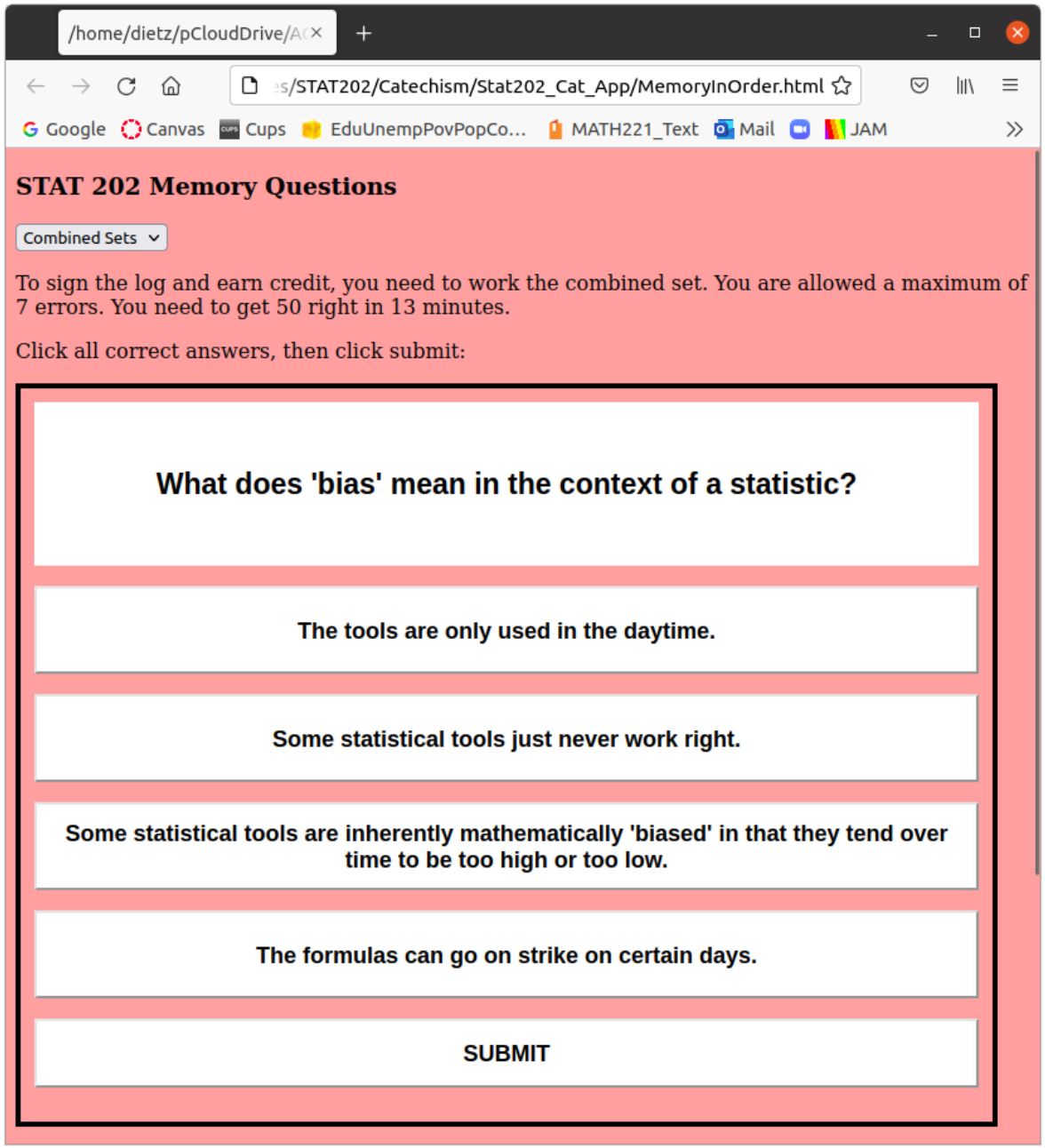

/home/dietz/pCloudDrive/A
s/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
Let's say we group all current AU students into grads and undergrads, and first, we want to know the ratio of how many grads are wearing jeans today.
Total number of grads OVER total number of grads wearing jeans.
Total of all students wearing jeans OVER total number of all students.
Total number of grads wearing jeans OVER total number of grads.
Total number of grads OVER total number of students (grads plus undergrads).
SUBMIT

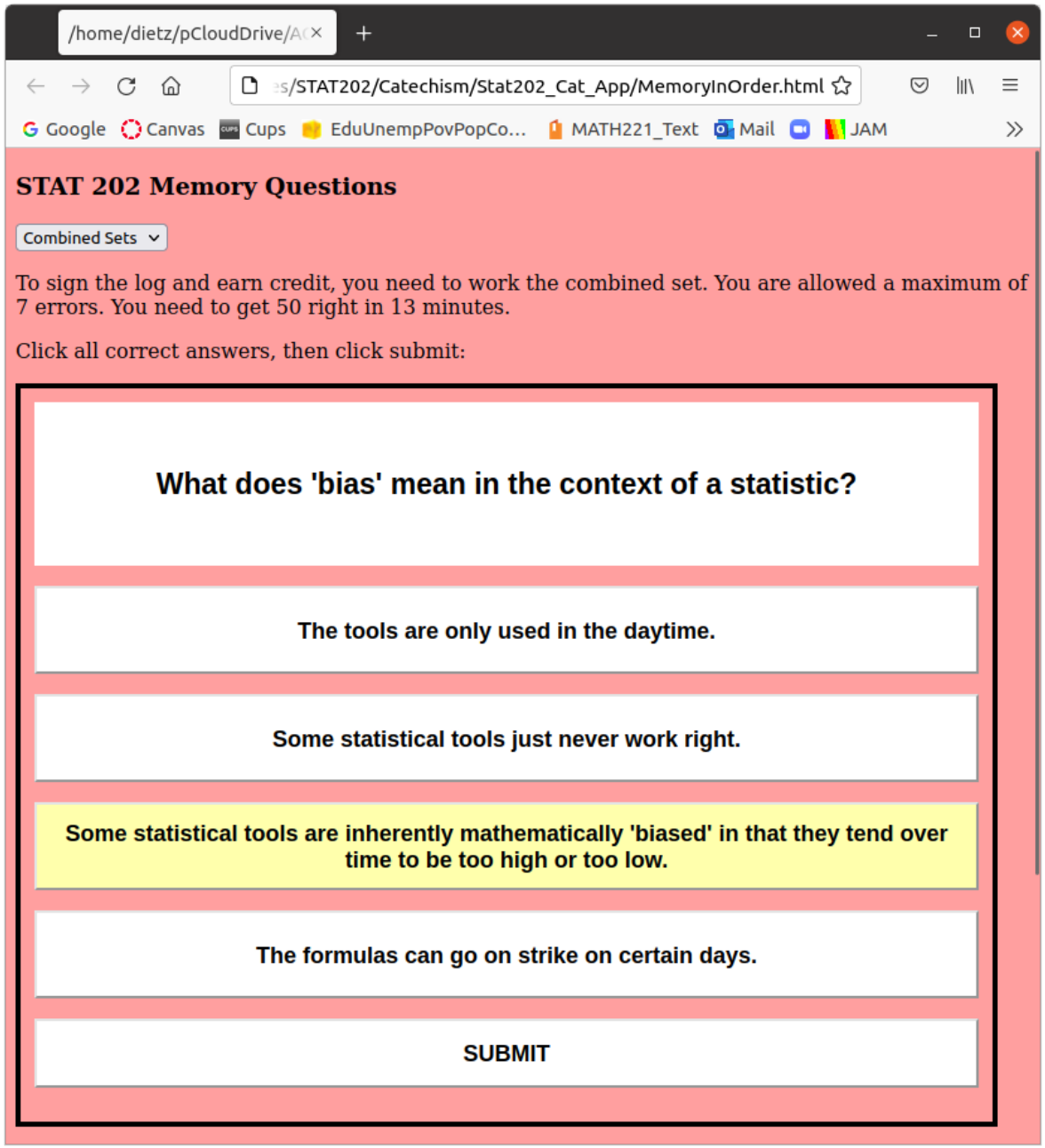

/home/dietz/pCloudDrive/A
s/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
What is an expected value in the context of a 2-way contingency table?
Row sum PLUS column sum over total sum.
Actual number with both traits OVER overall total in all cells.
Row sum TIMES column sum divided by overall sum.
Percent with column trait TIMES percent with row trait TIMES total number overall.
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 an expected value in the context of a 2-way contingency table?
Row sum PLUS column sum over total sum.
Actual number with both traits OVER overall total in all cells.
Row sum TIMES column sum divided by overall sum.
Percent with column trait TIMES percent with row trait TIMES total number overall.
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 can cause 'bias' in the context of a survey?
The only place the survey is given is in front of the gym.
If the person asking the questions is wearing a political button
The surveys are only done in the evening.
A survey is written in such a way as to encourage a certain response.
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 can cause 'bias' in the context of a survey?
The only place the survey is given is in front of the gym.
If the person asking the questions is wearing a political button
The surveys are only done in the evening.
A survey is written in such a way as to encourage a certain response.
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 does 'bias' mean in the context of a statistic?
The tools are only used in the daytime.
Some statistical tools just never work right.
Some statistical tools are inherently mathematically 'biased' in that they tend over time to be too high or too low.
The formulas can go on strike on certain days.
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 does 'bias' mean in the context of a statistic?
The tools are only used in the daytime.
Some statistical tools just never work right.
Some statistical tools are inherently mathematically 'biased' in that they tend over time to be too high or too low.
The formulas can go on strike on certain days.
SUBMIT