$Q - Q$ Plots

Donna Dietz

American University

dietz@american.edu

STAT 202 - Spring 2020

What is it?

Definition

Q stands for Quantile. A $Q - Q$ plot is short for "Quantile $-$ Quantile plot.

So what does that mean?

First

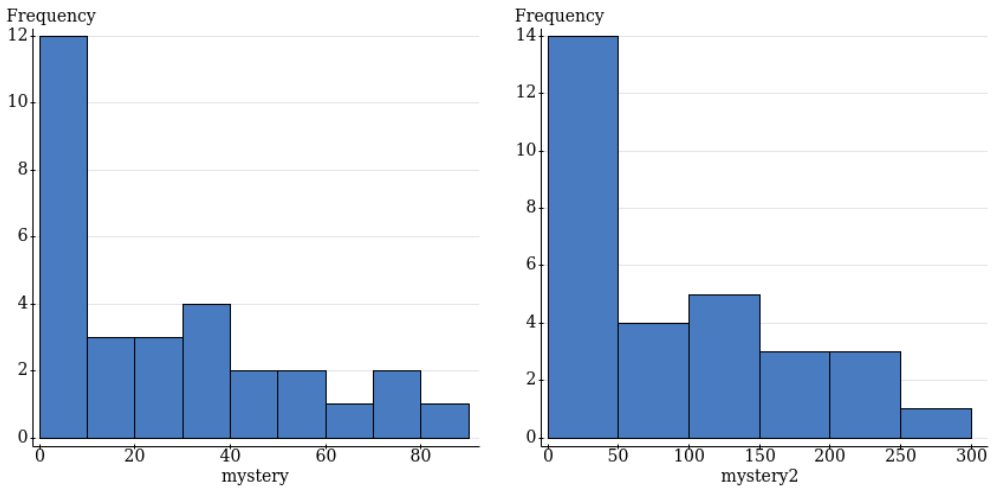
Before we tear apart this definition, let's take a high-level view of what this means.

Objective

The objective of this trick is to see if two sets of data or if two distributions are **roughly the same shape**.

Example

For example, I have two sets of data, and I made histograms for each set. They look similar. Are they similar?



The Question

The question is really whether or not there's a function

$$y = mx + b$$

that can take you roughly from one set of data to the other.

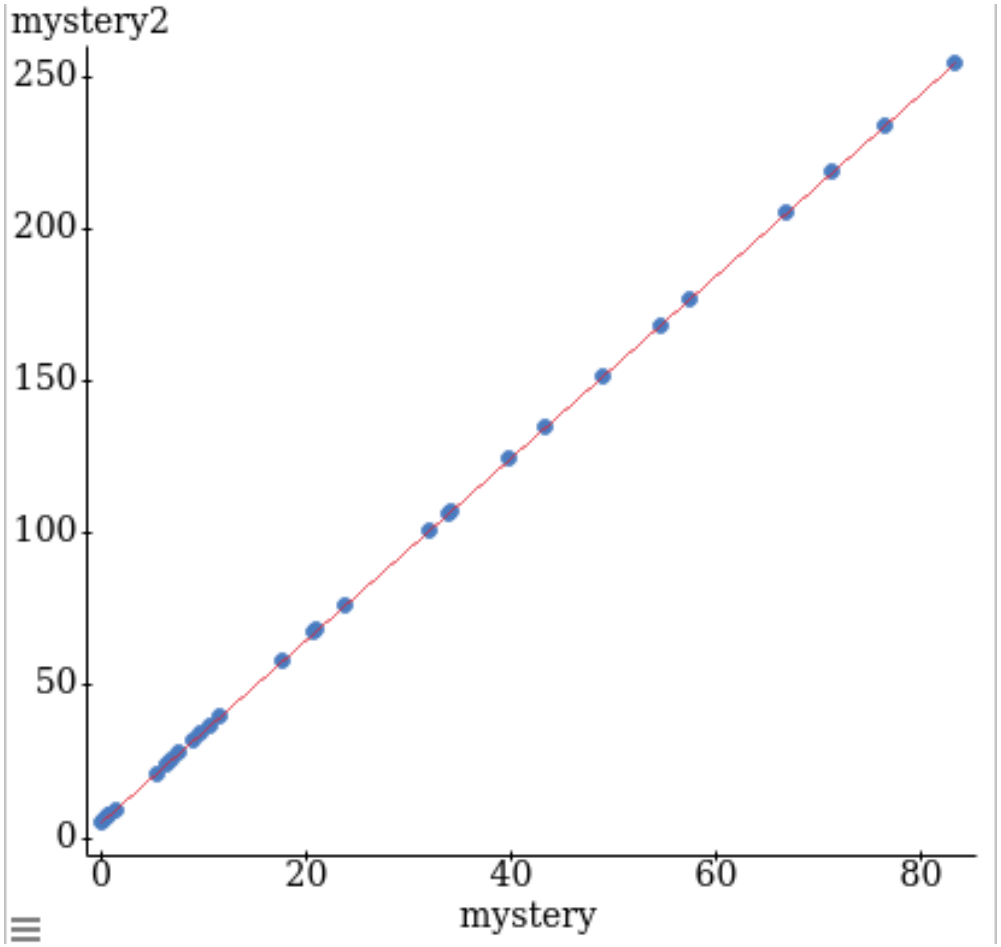
In order to tell if this is the case, if they happen to be the same exact size, the easy way to do this is to just make a scatterplot of the sorted data.

So...

So, if I had a sorted data set: $x = \{2, 4, 5, 6\}$
and another sorted data set: $y = \{24, 28, 30, 32\}$

Where we happen to know secretly that $y = 2x + 20$, but we are pretending we didn't know that yet. We can just do a scatterplot with a best-fit line. If a line is a good fit, this is similar to what you already know about best-fit lines, with a few extra add-on tips.

So plot it!



You can see right away that this was a setup! The x values and y values are exactly related by a line. It's not usually this perfect!

What perfection looks like:

Simple linear regression results:
Dependent Variable: mystery2
Independent Variable: mystery
mystery2 = 5 + 3 mystery
Sample size: 30
R (correlation coefficient) = 1
R-sq = 1
Estimate of error standard deviation: NaN

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	5	NaN	≠ 0	28	NaN	NaN
Slope	3	NaN	≠ 0	28	NaN	NaN

If you do a LSR on points which lie precisely on a line together, this is what your LSR report will look like. Notice that $R = 1$, so of course 100% of the variation is explained by the model, which is $\mathit{mystery2} = 5 + 3 \cdot \mathit{mystery}$.

Also, notice the *NaN's* all over the Parameter Estimates table. This happened, because we had a lot of division by zero going on.

Why sort?

Why do the data need to be sorted first?

As you may have expected, the word "Quantiles" is important. For example, "Quartiles" are a type of quantile where $n = 4$. So, if we were doing a *quartile-quartile* plot, or if we were lining up *five number summaries*, we would check to see if we had a line that could take us from $Q1$ of x to $Q1$ of y and the same line to take us from the median or $Q3$ of x to the same location in y.

What if the data are of different sizes?

So, if the data are different sizes, we really have to go to a different level of mathematics to discuss a good way to do this. We can easily ask StatCrunch to do it for us, but it's not easy to explain or to do by hand.

However, there is another case that's easy to explain or even do by hand, and that's the case when you wish to compare your data against a standard normal curve, and that is a very common thing you would want to do. Officially, these are called *Normal-Quantile Plots* or sometimes just *normal plots*.

Normal-Quantile Plots

In this special (and very common) case, we wish to compare our data against the quantiles of a typical normal curve. Let's take as an example what happens if our data set has just three values:

$$S = \{5, 10, 15\}.$$

What type of quantile has three elements? Quartiles do. We have

$$N = \{Q1, Q2, Q3\}.$$

So the question is, if you look at a Standard Normal Table, where would you expect $Q1$, $Q2$, and $Q3$ to be?

Finding $z-scores$

Looking at the table, we want to see where 0.2500, 0.5000, and 0.7500 land.

The middle one is the easiest. Half the data fall to the left of $z = 0$.

To find the $z-score$ for $Q1$, look for the closest match to 0.2500 in the table. We fall between two values, pretty close to the middle right between them. For $z = -0.67$ the area to the left is 0.2514 but for $z = -0.68$, the area to the left is 0.2483, so we can use $z = -0.675$ as the $z-score$ for $Q1$.

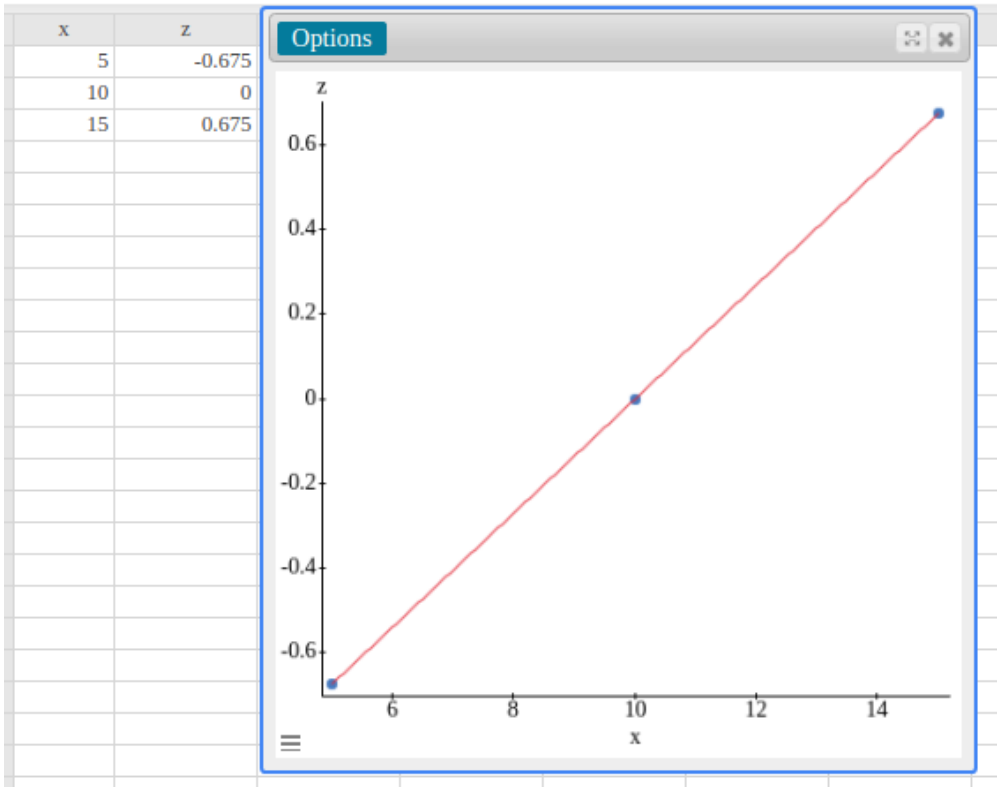
By symmetery, we can use $z = 0.675$ for $Q3$.

Standard Normal probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483

Pointless plot

Although this is not really going to give us useful information, let's go ahead and plot the x values against the z values we found.



In this case, the points were either equally spaced or weren't. There is not much more to analyse about three sorted points.

Realize...

Realize that the $z - values$ you select for the $Normal - Quantile$ plots are the same for every data set of the same size.

What about 7 points?

For seven points, you can calculate the $z-scores$ or Octiles. This is convenient for us, because we just did the Quartiles so we really only need to find two more values on the table. Symmetery gives us the remaining two values.

For an area of 0.1250, we will take the closest match, or 0.1251 in the table, for a $z-score$ of $z=-1.15$. For an area of 0.3750 we can accept 0.3745 with a $z-score$ of $z=-0.35$.

Combining this with the $z-scores$ we found for the quartiles, we get the $z-scores$ for the seven octiles as:

$$z=\{-1.15,-0.675,-0.35,0,0.35,0.675,1.15\}$$

So, any data set with 7 items can be sorted and plotted against these $z-scores$ so that the data itself can be compared against a standard normal distribution.

Generically, for n points

For n points, you would look for quantiles with n points, or just add one to n. Recall, 3 points gaves us quartiles. Nine points would give deciles. Ninety-nine points would give us percentiles.

Where n is the number of points, the first area you would look for in the table is $\frac{1}{n+1}$. So, for 99 points, you would first look for $1/100 = 0.01$ in the table, which is close to $z = -2.33$.

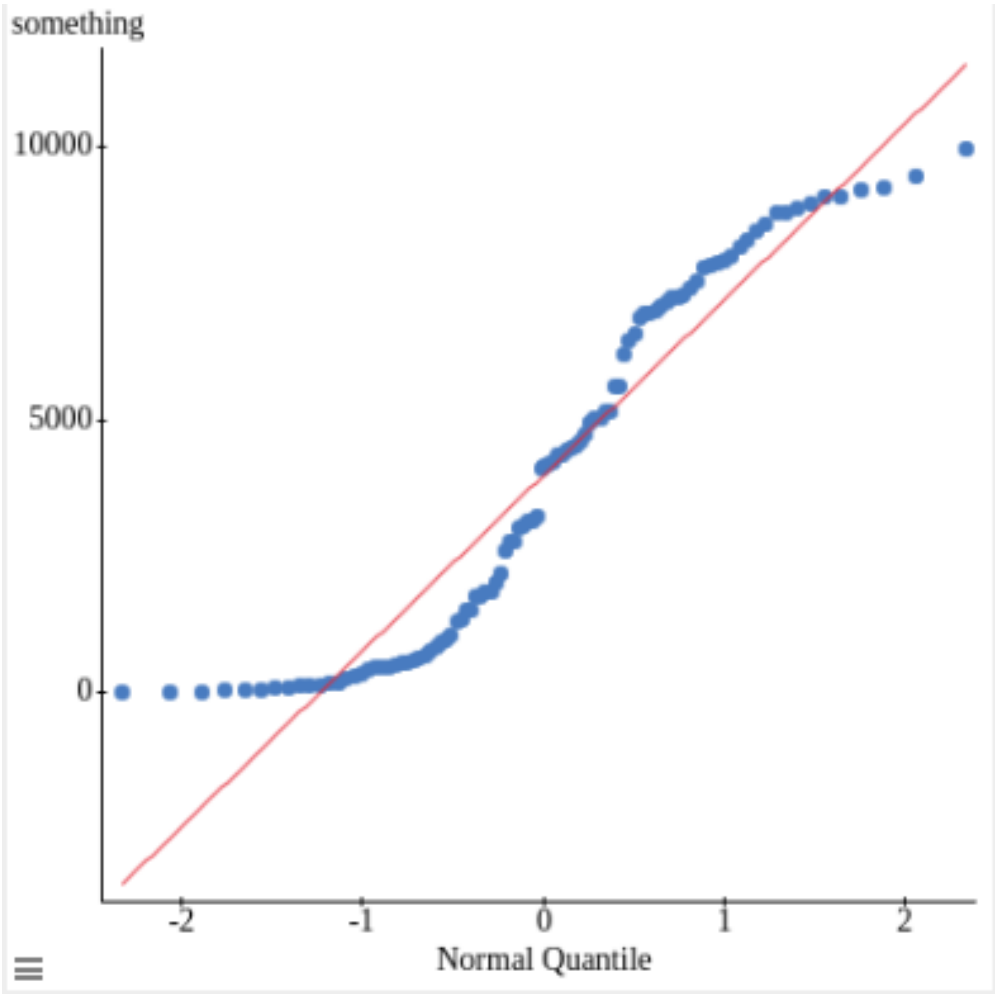
Continue looking for $\frac{2}{n+1}$, $\frac{3}{n+1}$, $\frac{4}{n+1}$ and so forth until $\frac{n}{n+1}$.

Luckily for us, this is annoying for humans but easy for computers to do.

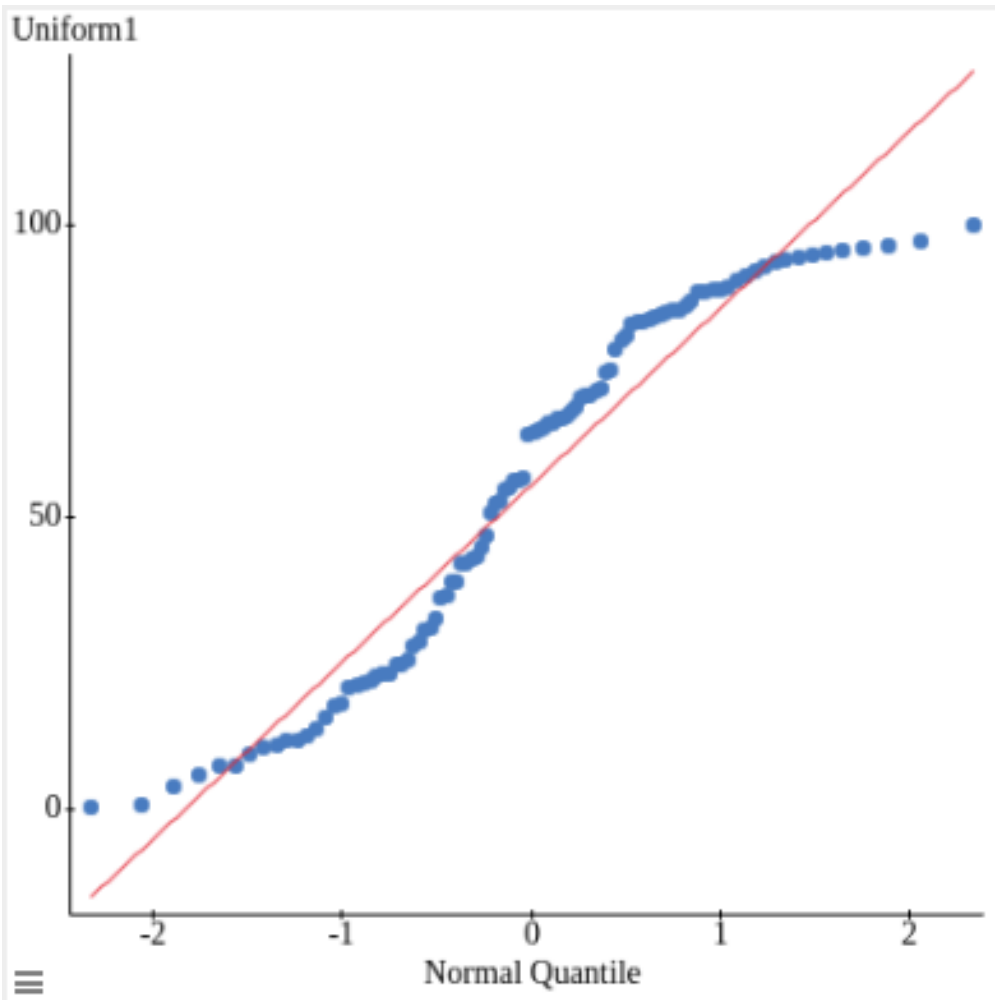
How this can go wrong

We had a similar thing going on when we looked as the residuals. We are looking for weird non-line things crossing over/through the line of best fit. Let's compare two distributions that are definitely not the same and see what happens.

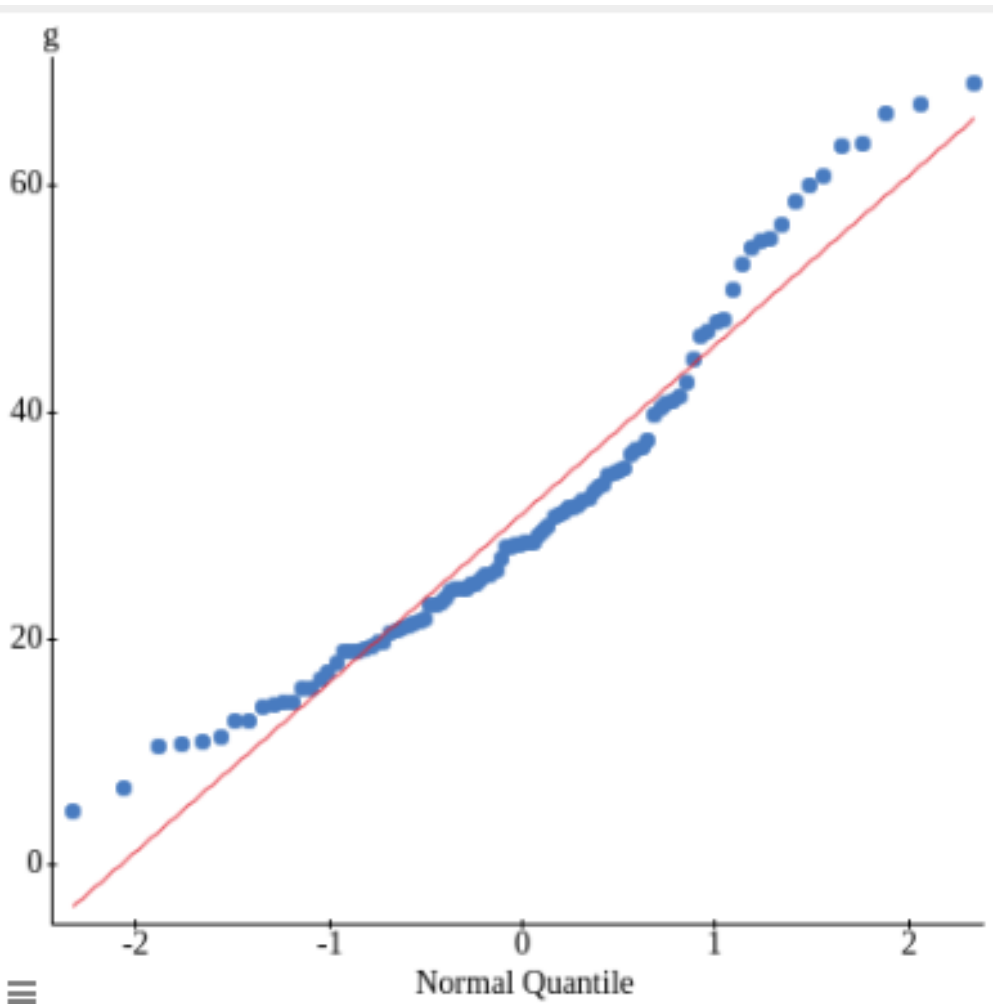
A bad one



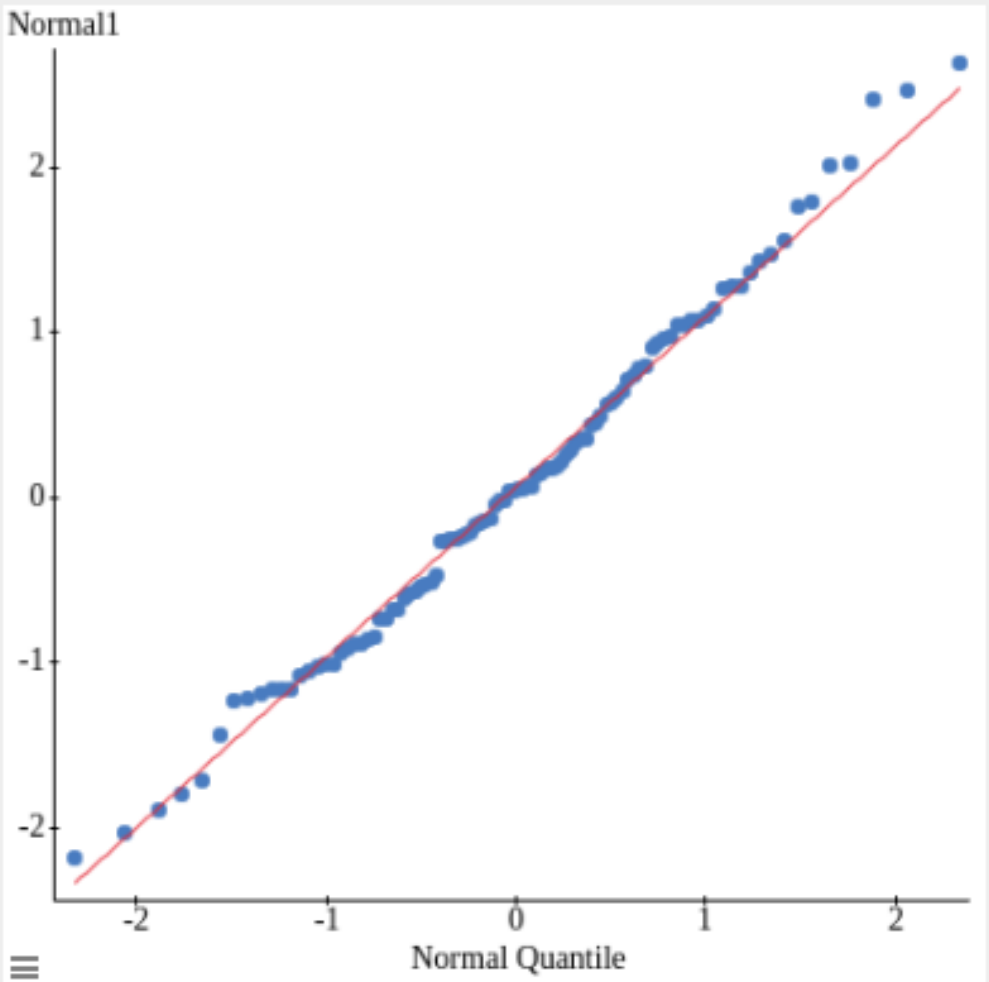
Another bad one



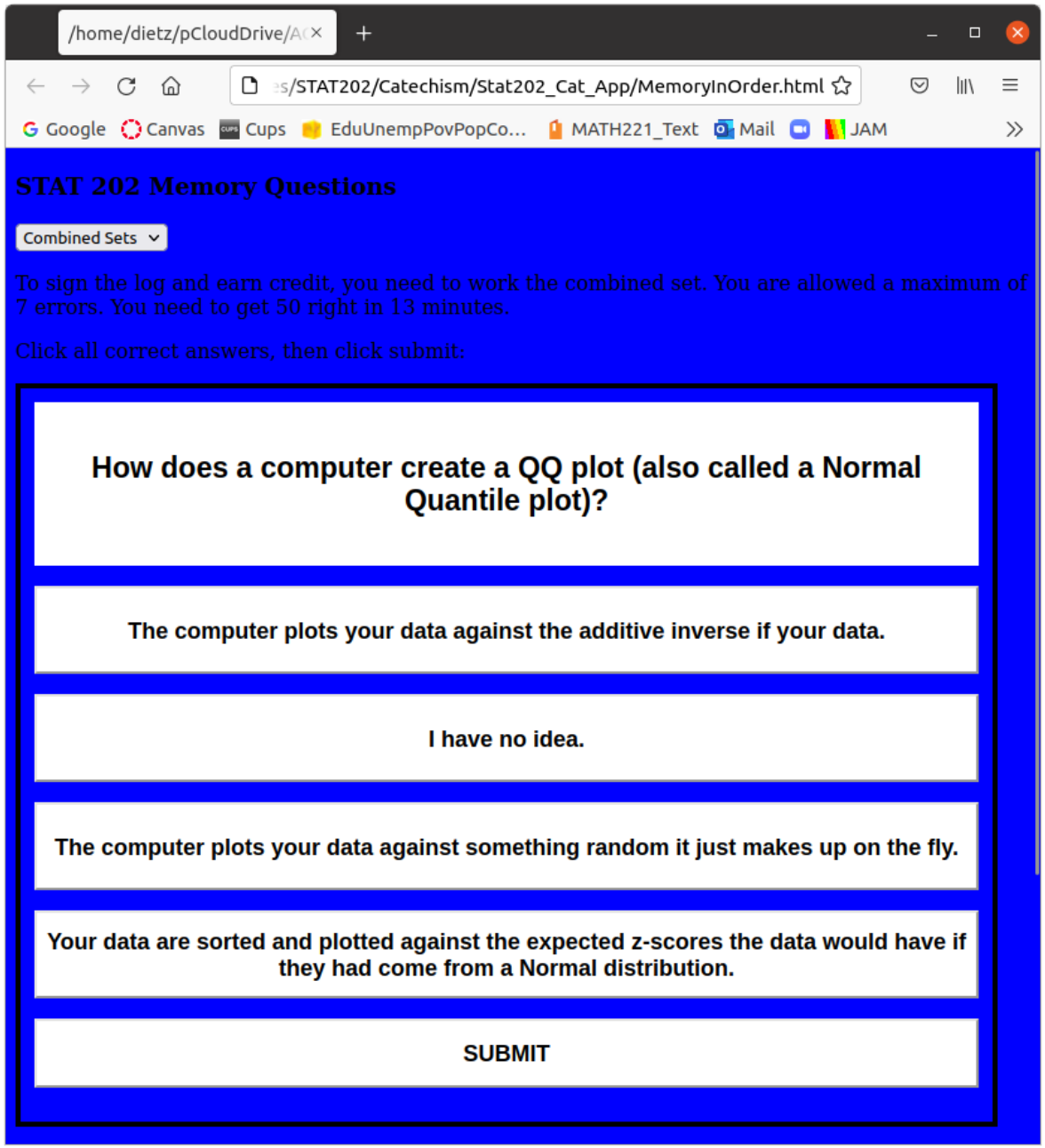
Another bad one

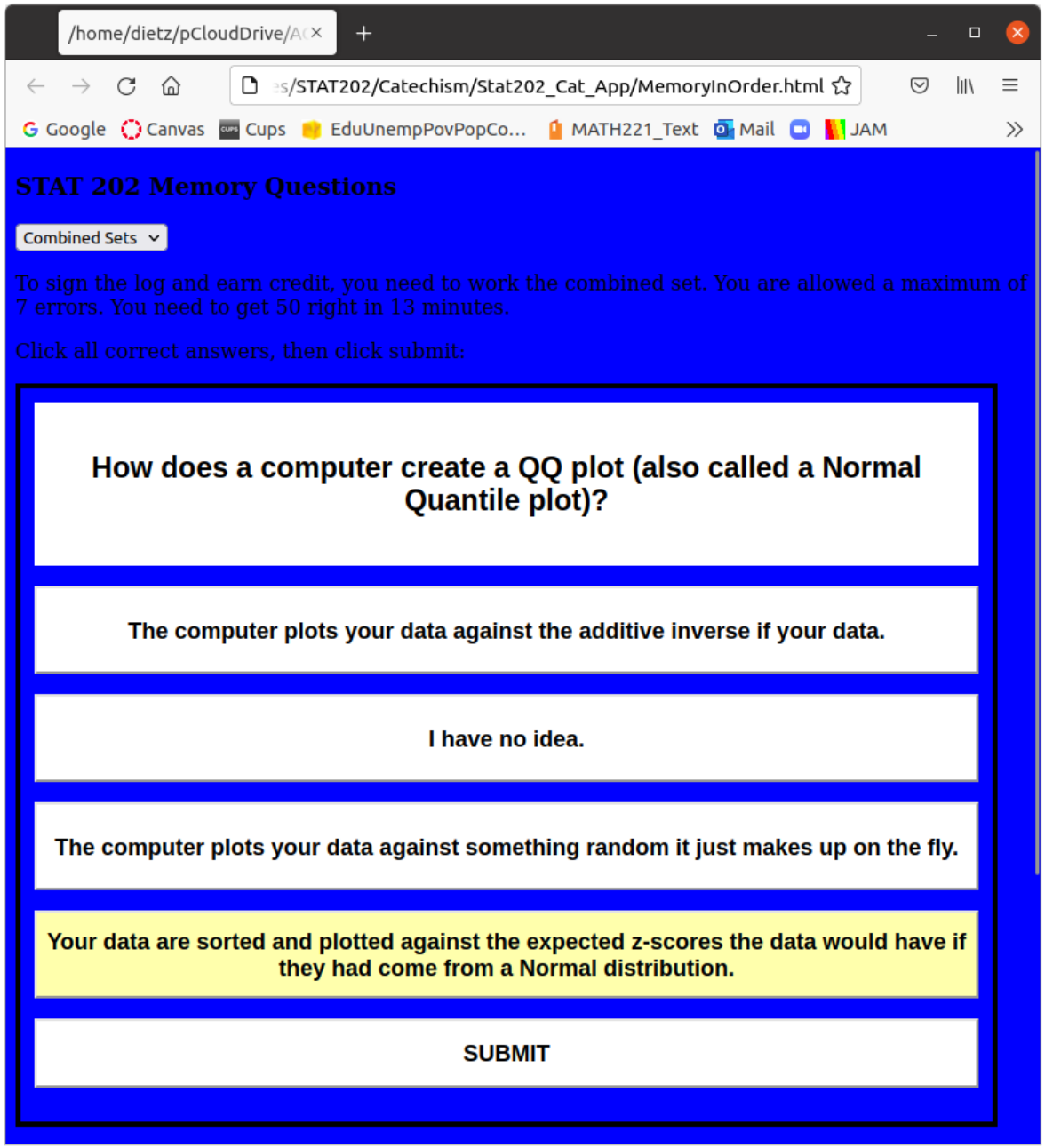


A good one



MEMORY QUESTION


/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:
How does a computer create a QQ plot (also called a Normal Quantile plot)?
The computer plots your data against the additive inverse if your data.
I have no idea.
The computer plots your data against something random it just makes up on the fly.
Your data are sorted and plotted against the expected z-scores the data would have if they had come from a Normal 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:
How does a computer create a QQ plot (also called a Normal Quantile plot)?
The computer plots your data against the additive inverse if your data.
I have no idea.
The computer plots your data against something random it just makes up on the fly.
Your data are sorted and plotted against the expected z-scores the data would have if they had come from a Normal distribution.
SUBMIT