Answering Questions With Government Transparency Data

An Exploration of private sector jobs

Jared Anderson

August 2, 2023

My project is largely an exploration of Employee data for all government employees in Oregon from 2015 to 2022. (Data found at https://www.oregon.gov/transparency/Pages/State-Salaries.aspx)

There are some questions that I believe I could answer using this data, some more relevant to me than others.

0.1 Some Assumptions and Caveats

1. This data is real life and economic in nature, which is to say, messy and somewhat unnatural
    - I will be using an α of 0.05–this isn't physics. As this is essentially economic data, this seems appropriate.
    - Confidence intervals will be set to 90% in keeping with our generosity in terms of precision.
    - When calculating certain statistics, I will be sampling from the data. This is to allow certain statistics to be computed using the tools I have learned. Other times this is done to estimate a simpler distribution than the entire population's real distribution. Any time this is done, it will be explicitly stated.
2. There is a large amount of data, (approximately 27MB of text data). As such, I will be making extensive use of R to calculate certain values and statistics.
    - I will be displaying only partial and/or relevant results in tables.
    - Code for the calculations will be made available at https://github.com/Laharah/ma255stats
3. This is for class. I'm trying to show I learned anything, sometimes the data makes that hard to do:
    - There will be times where in order to apply a particular statistical test, I will make the assumption that the data is normally distributed. This may not always be the case. Whenever this assumption is made, I will say so explicitly.
    - When feasible, the first time a statistical test is encountered I will be showing how the result of a statistic is calculated. Other statistics will be summarized. I will also be briefly summarizing the methods I used to filter and manipulate the data. Again see the code listing to see how a particular calculation was done.

1 An Exercise In Conditional Probability

I met a woman at a python conference in early 2020. She said she worked or had worked for the government as an analyst—I can't remember which. We talked about what we did for a while, and at one point I remember asking how well her job payed; she said that she made "around 65 grand" per year.

Looking at the data, there are analysts that work in over 50 different departments. I'm curious which department she was most most likely working at. This will let me get a "feel" for working with the data, and offers a good opportunity to practice. But first:

Assumptions & Methods:

- I will be assuming that analyst pay in each dept is normally distributed.
- I will be sampling from the dataset to estimate these distributions.
- I will be disregarding departments that have less than 10 analysts to sample from.

1.1 Filtering And Sampling The Data

My plan for sampling is to do as if I had surveyed 10 analysts from each department that employs analysts and recorded their department and their annual salary. These samples will be random both by individuals and by time in the range from 2015 to 2020.

1.2 Is Sampling By Department Necessary?

If every department pays it's analysts the same average wage, there may be no reason to sample from each department individually. To test this, I'm going to conduct an **Analysis of Variance** or ANOVA test.

For this test I'm going to assume:

1. Each department's salary distribution is normal (as was stated above).
2. Each department has the same standard deviation (σ) in the amount they pay their analysts.

These two assumptions are required to conduct an accurate ANOVA test. The other assumption requred for an ANOVA test is that the samples are simple random samples, which I have ensured. I will point out that it is unlikely that each departments pay has the same σ^2. However we can hope that they are similar enough to get some kind of useful information out of the test.

The groups for the ANOVA test are going to be the departments and the values will be the samples taken from each department.

Our hypotheses are as such:

$H_0 : \forall d \in \Omega \quad \mu_d = \mu_\Omega$, that is: each department pays their analysts the same on average.

$H_A : \exists d \in \Omega$ where $\mu_d \neq \mu_\Omega$ or, **not** every department pays their analysts the same on average.

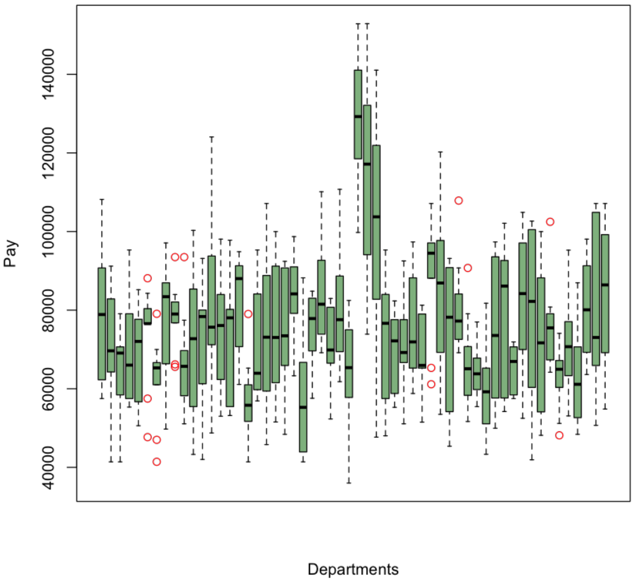
This results in the following ANOVA table:

Table 1: ANOVA of Dept. Analyst Pay

	Df	Sum Sq	Mean Sq	F value	P-value
dept	55	79328877331	1442343224	6.1266	2.7597e-30
Residuals	504	118653001472	235422622		

This P-value is extremely small, giving us high confidence that the departments **do not** pay their analysts the same on average. However given our assumption of normality and equal variance, we should check our work visually:

Figure 1: Boxplot of Analysts Pay



It's pretty clear looking at the chart that there is quite a lot of variance in the pay among the departments, With a few particular outliers in the middle.

1.3 Bayesian Analysis

Now, to calculate the probability of the woman I spoke to working in any given department, I must calculate the *conditional probability* that she works there. In other words, We will calculate the probability that she works at a particular department *given* that she is paid a certain amount. This is written as $Pr\{\text{working in dept. } x|\text{she is paid } y\}$

The formula to find this is called Bayes formula and it is written:

$$Pr\left\{A|B\right\} = \frac{P(B|A)P(A)}{P(B)}.$$

Using these probabilities as analogs for elements of our problem, we assign them like so:

$$\begin{aligned} P(A) &= \text{The probability that she works at a particular department} \\ P(B|A) &= \text{The probability that she makes gets a certain salary } \textit{assuming} \text{ that we know} \\ &\quad\ \text{she works at a particular department} \\ P(B) &= \text{The } \textbf{total} \text{ probability that she makes a particular salary} \\ Pr\left\{A|B\right\} &= \text{The probability that she works at a particular department } \textit{given} \text{ that we know} \\ &\quad\ \text{what her salary is. This is the value we're looking for} \end{aligned}$$

Now I'll cover *how* you would extract these number from the data.

- $P(A)$: We can calculate this by taking a department, counting the **total** number of employees there, and dividing by the total of **all** the employees.
- $P(B|A)$: Since we've assumed that all departments pay their analysts with a normal distribution, we can use our samples to estimate the average and standard deviation of a particular department's payscale. Using this estimated normal distribution, we can calculate the probabilty that she makes a certain amount of money (within an arbitrary but reasonable range) assuming that she worked there.
- $P(B)$: To calculate this we must *sum* $P(B|A)$ across **all** departments.

Solving the above equation for every department, we can find the Departments with the highest probability that she worked there. The only variable that I need to supply is the range of salaries we should be testing for. I decided (again arbitrarily) that given she said she makes about $65,000, a good lower range is about $63,000, any less and she would probably have said she makes about $60k. Likewise I set the upper bound of the range to $67,000, any more and I think it likely she would have told me she makes about $70k.

Here is a look at our most significant results.

Table 2: Most Likely Depts. My Acquaintance Worked

Department	$Pr\{Dept.|sal.\}$
ADMINISTRATIVE SRVCS, DEPT OF	0.0356
CONSUMER AND BUS SRVCS, DEPT	0.0364
PUBLIC EMPS RETIREMENT SYSTEM	0.0617
TRANSPORTATION, DEPT OF	0.1032
OREGON HEALTH AUTHORITY	0.1821
HUMAN SERVICES, DEPARTMENT OF	0.2185

looking at the table, we can conclude that It is most likely that she was working in the Department of Human Services, the Oregon Heath Authority, or the Department of Transportation (in that order)

As a sanity check, we can check an make sure that the sum of all these probabilities is equal to one-which I am happy to report is the case.

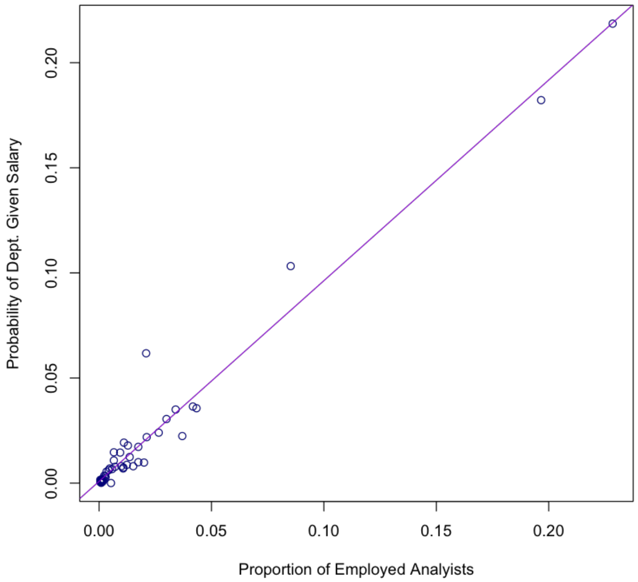
1.4 Was This A Waste of Time?

Having been looking at this data for a while I couldn't help but notice that the top 3 most likely departments are also the departments that employ the most analysts.

This makes me wonder if we could have simply found the departments that employed the most analysts and been just as well off.

To check, I'll graph our calculated probabilities against the proportion of all analysts each department employs and calculate the correlation coefficient.

Figure 2: How much Time Did I Waste?



After looking at the graph, it should not surprise you to hear that the calculated correlation coefficient, $r = 0.9829$; an almost perfect one-to-one relationship. We could have saved a ton of work if we had just looked at the porportions of total analysts for each department. It should be noted that the most likely Departments (Human Services, The Health Authority, and Dept. of Transportation), are true outliers. The lower section of the graph looks much more interesting. If we somehow knew she *didn't* work in those departments, we could eliminate them from our calculations and hopefully get more interesting results.

2 Should I Get a Job in Government?

I have been working for myself as an IT contractor for more than a decade now. While I like the pay and the near total flexibility it affords me, there are certain charms to the idea of having a "normal" job. Structure, a matching retirement account, and above all else: sweet sweet healthcare that you don't have to pay for.

The pros and cons are difficult to weigh against each other-there are a lot of factors. The next few questions will concern answering questions that could help me make such a decision.

2.1 Is The Oregon Public Sector Even Hiring?

How difficult would it be for me to get a job in government-any job? To help me answer this question, I can analyse the data to see if Oregon is likely open more positions over time. Also, I could use this information to calculate how many jobs I can expect them to add this year.

To calculate this statistic, we will first gather the entire dataset (our whole "population") and organize it by year and then count the total number of employees for that given year.

Once that's been done (not as easy as it sounds when you don't really know R), we get a set of values that looks like this:

Table 3: Total Workers Per Year

Year	2015	2016	2017	2018	2019	2020	2021	2023
Workers	36767	37064	33219	37187	38033	39160	38740	45068

Now that we have these values it's fairly trivial to get the mean and standard deviation for both rows of data:

$$\begin{aligned}\overline{x} &= 2018.5\\ \overline{y} &= 38154.875\\ s_x &= 2.4495\\ s_y &= 3329.0208\\ &.\end{aligned}$$

Using these values I can now calculate the **correlation coefficient**, that I somewhat glossed over above. The correlation coefficient is denoted by the letter r, and it's equation for a given set of points is:

$$r_{x,y} = \frac{1}{n-1}\left[\left(\frac{x_1-\mu_x}{\sigma_x}\right)\left(\frac{y_1-\mu_y}{\sigma_y}\right)+\left(\frac{(x_2-\mu_x)}{\sigma_x}\right)\left(\frac{y_2-\mu_y}{\sigma_y}\right)+\dots\right].$$

If we plug in the values above and solve, we get $r = 0.7460$, which indicates a somewhat strong correlation between the year and the number of government employees. Using this value we can then calculate the least squares linear regression line of the data.

The formula of the LSLR line is

$$\hat{y} = \beta_0 + \beta_1 x.$$

where β_1 is the average rate of change in the number of workers per year, and β_0 is the y intercept of the data (which in our case is somewhat nonsense, as there is no year 0 A.D, and if there was, Oregon wouldn't have existed, and it certainly couldn't have had the millions of "negative" employees the model predicts it to have had).

Notice that we don't have to estimate this regression line, because we have access to the actual values from the entire population. The coefficients (β_0 and β_1) can be calculated as such:

$$\beta_1 = r\frac{\sigma_y}{\sigma_x} \tag{1}$$

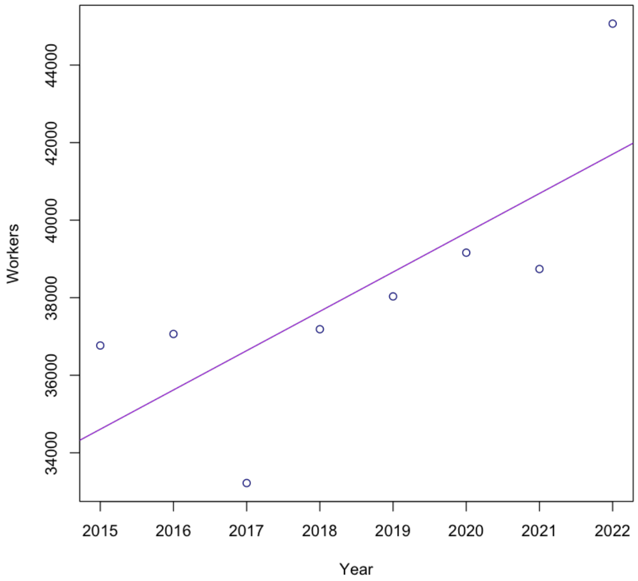
$$\beta_0 = \mu_y - \beta_1\mu_x \tag{2}$$

Plugging in our numbers from above and solving gives us:

$$\hat{y} = 1014x - 2008196.$$

This equation will allow us to graph both our points and the least squares regression line like so:

Figure 3: Yearly Oregon Government Job Growth



Lets see how confident we can be that we've gotten the slope of the regression line by calculating a **confidence interval**. We'll need two things:

1. The standard error of β_1:
$$\sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$
2. and the quantile of the student's-t distribution with degrees of freedom d over the interval $[\alpha/2, 1-\alpha/2]$:
$$t_{d,\alpha/2}.$$

The standard error has a term s_e in it that hasn't been calculated yet. It's formula is:

$$s_e = s_{y|x} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

I leave the calculation of SE_{β_1} as an exercise for the reader.

Plugging in our values and getting the t-distribution quantiles from a computer or a table we get

$$SE_{\beta_1} = 369.5217 \quad (3)$$
$$t_{7,0.25} = 1.9432 \quad (4)$$
$$(5)$$

Now we can finally calculate our confidence interval:

$$\begin{aligned} \text{C.I} &= \beta_1 \pm t_{n-2,\alpha/2} \times SE_{\beta_1} \\ &= [295.7503, 1731.845] \end{aligned}$$

.

which to be frank is not great for a 90% confidence interval. It seems we have a bit of a precision problem. That said, let's test our hypothesis that the number of employees is rising year-over-year *at all*. To be more formal let's state our null and alternative hypotheses.

$$\begin{aligned} H_0 : \beta_1 \leq 0 &\quad \text{or there is a negative or no difference in employees each year} \\ H_A : \beta_1 > 0 &\quad \text{or there is a positive correlation between the year and the number of employees.} \end{aligned}$$

To calculate the p-value we'll calculate a t-stat first.

$$\begin{aligned} t\text{-stat} &= \frac{\beta_1 - m_0}{SE_{\beta_1}} && (6) \\ &= 2.7435 && (7) \\ & && (8) \end{aligned}$$

And the last thing we need is to use the t-distribution to get our p-value:

$$\begin{aligned} P\text{-Value} &= Pr\left\{t_{n-2} > t\text{-stat}|H_0\right\} \\ &= 0.0168 \end{aligned}$$

.

As I said in the beginning, we've given ourselves a generous α of 0.05. This P-value gets us well under that line, which means that we can confidently **reject** the null hypothesis.

Technically I've answered my question, I'm confident that the number of government employees is growing. But I want to be a little more specific; if I apply for a position this year, do I expect there to be a job there for me?

Answering this is not only straight forward, but easy. All I have to do is plug in the desired year into my least square linear regression model:

$$\begin{aligned} \hat{y} &= \lfloor \beta_0 + \beta_1(2023) \rfloor \\ &= -2351 \end{aligned}$$

.

It seems that the answer to that question is a resounding no—my model predicts that Oregon will lay off more than 2000 people next year. It seems like the reason for this result is that 2023 was a noticeable positive outlier in terms of number of employees. My model expects there to be some kind of regression to the mean. But I suppose it's possible that there is a factor or variable that can't be found in my data, and 2022 marks the beginning of a new growth period.

2.2 How Many New Positions That I Might Be Interested In Do I Expect To Be Added Next Year?

First I will filter out jobs that I'm not particularly interested in doing. This leaves me with a dataset of 6068 jobs over the last eight years. Of these jobs, how many do I expect will have openings this year?

To answer this question we won't really be doing anything new. We'll be doing more or less the same as we did in section 2.1. First we'll group together each remaining job by year and calculate a least square regression line. A quick summary of the data and variables is listed here:

Table 4: Candidate Jobs Per Year

Year	2015	2016	2017	2018	2019	2020	2021	2023
Jobs	741	740	707	739	781	773	789	798

$$\begin{aligned} r &= 0.8218 \\ \beta_1 &= 10.524 \\ \beta_0 &= -20483.810 \\ SE_{\beta_1} &= 2.979 \\ t\text{-value} &= 3.533 \end{aligned}$$

.

Frankly I'm more interested in what I can roughly expect than I am about the surety that my trend line is positive, so I'll calculate a 90% confidence interval. Honestly that's more confident than I've been about almost anything in real life lately which makes it good enough for me:

$$\begin{aligned} \text{C.I.} &= \beta_1 \pm t_{6,0.05/2} \times SE_{\beta_1} \\ &= [4.7349, 16.313] \end{aligned}$$

.

That gives me some hope, it certainly seems likely that there is slow but steady growth in my candidate jobs. Let's take a look at the graph:

Now to see how many openings total I should expect this year:

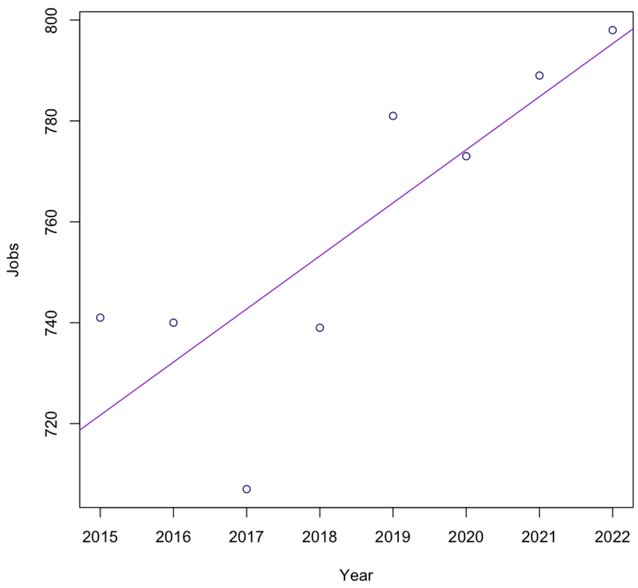
$$\begin{aligned} \hat{y} &= \lfloor 10.524(2023) - 20483.810 \rfloor \\ &= -2 \end{aligned}$$

.

Once again, I am likely out of luck if I'm looking to switch into a government job in the immediate future.

H

Figure 4: Candidate Jobs Per Year



2.3 Which of The Specific Jobs I'm Willing To Do Are Growing?

While I may be fairly confident that there is *overall* growth in my candidate jobs, which specifically are those? If they're just jobs that I can tolerate, without much window for growth, I may not be terribly interested.

To answer this question, I'm going to separate the jobs not only by year, but also by job type. This will give us 23 jobs with 8 years of records each.

I'll then take each job type, and calculate a least square linear regression on the number of employees in that job type based on the year. I'll filter this data *again* by ignoring any regressions that have negative growth but also any jobs where I'm not confident enough that my calculated growth isn't a mirage (where the P-value is too high.)

Though I already expect the answer, I'll also calculate how many openings in those jobs I can expect this year.

All of which leaves me with this:

Table 5: Most Likely Career I Can Get

Career	Average Growth Rate	P-value	Estimated Openings in 2023
CUSTODIAN	1.5833	0.02628	-1
RESEARCH ANALYST	5.3809	0.03370	-18

Whelp, I'm glad that at least *something* has edged out custodian. It seems to have done so fairly well too in terms of growth. My model predicts much more research analysts to be let go next year. This is likely a function of the fact that there are many more researchers in the dataset, and that the number of custodians you need only goes up when you expand offices.

2.4 Which Jobs That I Could Do Have The Most Fair Pay?

looking at the data, there are many jobs descriptions that are followed by numbers. I take this to mean that you can be promoted from one level to another. If I took one of these jobs (again these are ones I'd be willing to actually do), it's important to me that I'm working with people that are being paid fairly—it makes for a less hostile work environment. In this case, I'll settle for the qualifier that the salaries in that job are normally distributed.

It's important that I be able to quantify a distribution's "normalness" so that I can rank one career against another. To do this I'm going to be using a Shapiro-Wilk test. Because the Shapiro-Wilk test tends to overfit with any sample size of about 60 or so, I'm going to sample from the population of each career. I'll take 30 simple random samples from each one to use in the test.

Before I do that though, I want to get an intuitive sense for how normal these distributions are. To do that I'm going to graph histograms and Q-Q plots for a few randomly selected careers.

A Q-Q plot is made by placing your observations in rank order ($y_1 \leq y_2 \leq \cdots \leq y_n$). We then shift these points slightly off center. Not by much, just enough so that we don't get strange values at the edges. We then scale these values by a factor of $1/n$. In more concrete terms, if j is the jth order of the numbers go through the function

$$f(j) = \frac{j - 1/2}{n}.$$

This pulls the numbers in so that $y_{(j)}$, roughly moves to it's quantile position so long as the distribution was normal to start with. These numbers become the normal scores or n-scores.

Once done, we can plot these n-scores against an actually standardized set of ordered *values*, $w_{(j)}$:

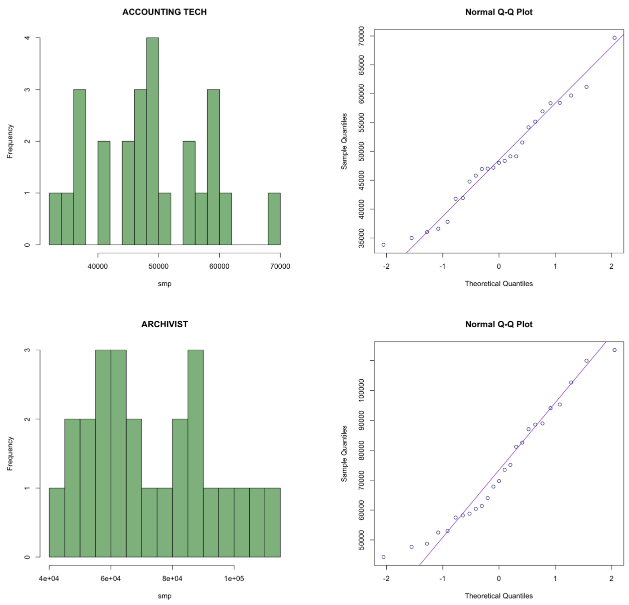
$$w_{(j)} = \frac{y_{(j)} - \overline{y}}{s_y} \approx z_{(j)}.$$

All of this allows you to plot the $w_{(j)}$ values against the "maybe" standardized values.

If the values were sampled from a roughly normal distribution, the graph should approximate a straight line along the $(1, 1)$ slope.

Here are a couple of the histograms and Q-Q plots for you to compare.

Figure 5: Sample Histograms and Their QQ-Plots



By looking at the graphs, you can see that the histograms are roughly normal (if you squint), but the Q-Q graphs are more clearly aligned along the 45 degree axis.

Finally we can put our samples through the Shapiro-Wilk test and take a look at the results:

Table 6: Careers and Their Salary "Fairness"

Career	Shapiro P-val
RESEARCH ANALYST	0.8094
ACCOUNTING TECH	0.6414
ASSOCIATE IN GEOLOGY	0.4636
TRUCK DRIVER	0.3448
ARCHIVIST	0.2627
TRANSPORTATION TELECOMMUNICATIONS SPECIALIST	0.0069

Well, it looks like **most** of the positions that I'm interested have strong evidence in favor of normality. Under my definition of egalitarianism, it looks like I'd be happy at any of these jobs except as a transportation telecom specialist (which is okay since I don't really know what that is.

2.5 Conclusion

I can't say that I've been completely swayed one way or the other as far as taking a government job is concerned. I certainly think that it's not something I should decide this year given my predictions from this data. I was surprised to see that most of the careers I'm interested in have "fair" pay.

To really help me decide, I should attempt to collect similar data for the private sector, match up like jobs, compare their salaries, growth, and "fairness".