

ONE QUANTITATIVE VARIABLE

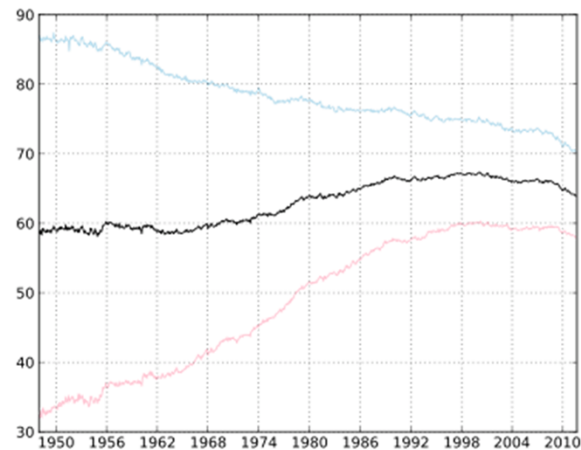
AND TESTS FOR PAIRED SAMPLES



Now let's review methods for one quantitative variable.

We will use an example which will result in a paired t-test regarding the labor force participation rate for women in the 60's and 70's.

Labor Force Participation Rate



The labor force participation rate (LFPR) is the proportion of individuals in a particular population who are currently working.

This graph (from http://en.wikipedia.org/wiki/Labor_force) shows:

In the center as a black line – the labor force participation rate for all US adults.

On the top – as a light blue line – the labor force participation rate for US men.

On the bottom – as a pink line – the labor force participation rate for US women.

There was a clear increasing trend among women and a decreasing trend among men over most of the period.

We are going to investigate data containing the labor force participation rate for women in 19 cities for two years, 1968 and 1972.

LFPR: Data $P(\text{participate in labor force} \mid \text{female})$

City	LFPR72	LFPR68
N.Y.	0.45	0.42
L.A.	0.50	0.50
Chicago	0.52	0.52
Philadelphia	0.45	0.45
Detroit	0.46	0.43
San Francisco	0.55	0.55
Boston	0.60	0.45
Pitt.	0.49	<u>0.34</u>
St. Louis	<u>0.35</u>	0.45
Connecticut	0.55	0.54

City	LFPR72	LFPR68
Wash., D.C.	0.52	0.42
Cinn.	0.53	0.51
Baltimore	0.57	0.49
Newark	0.53	0.54
Minn/St. Paul	0.59	0.50
Buffalo	<u>0.64</u>	0.58
Houston	0.50	0.49
Patterson	0.57	0.56
Dallas	<u>0.64</u>	<u>0.63</u>

Here is the raw data which comes from the data and story library.
<http://lib.stat.cmu.edu/DASL/Stories/WomenintheLaborForce.html>

Each pair represents the results for the listed city with the first value from 1972 and the second from 1968.

We will begin by summarizing the results in 1968 and 1972 individually.

In the data, the values corresponding to the min and max for each year are underlined.

You can see that the minimum in each year came from a different city. The maximum value in 1972 occurred twice and one of these cities – Dallas – was the maximum in both years.

Although not completely crucial to this problem. I will point out that these particular measures – the labor force participation rates – are actually estimates of a probability. In this case, we have a conditional probability. The labor force participation rate among women can be restated as the probability of a person participating in the labor force given the person is female.

We could write this in symbols as $P(\text{participate in labor force} \mid \text{female})$

I point this out mainly to illustrate that there are many applications of the concepts of probability that we discussed hidden in real-world problems in a wide variety of disciplines.

Labor Force Participation Rate

Variable	Minimum	Lower Quartile	Median	Upper Quartile	Maximum
LFPR72	0.350	0.490	0.530	0.570	0.640
LFPR68	0.340	0.450	0.500	0.540	0.630

Variable	Mean	Std Dev	Lower 95% CL for Mean	Upper 95% CL for Mean
LFPR72	0.527	0.071	0.493	0.561
LFPR68	0.493	0.068	0.460	0.526

Here we have the SAS output summarizing the labor force participation rate for these two years.

We can see that in 1972 all of the values for the 5-number summary are larger in 1972 than 1968 indicating an overall increase in this measure between these two years.

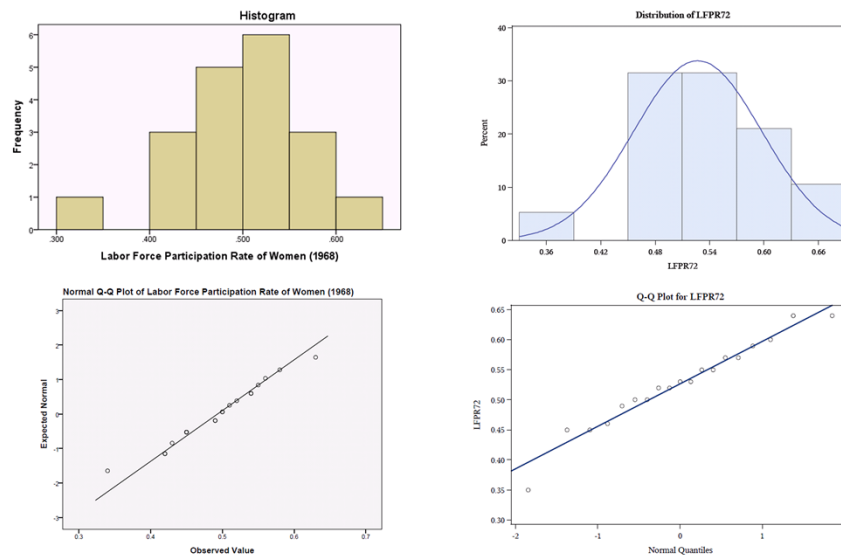
The mean LFPR in 1972 was 0.527 whereas in 1968 it was 0.493. Soon we will determine if this difference is statistically significant.

The variation within each year is similar as the range is exactly the same in both years, the standard deviation is similar (0.071 in 1972 and 0.068 in 1968), and the IQR is also similar ($0.57 - 0.49 = 0.08$ in 1972 and $0.54 - 0.45 = 0.09$ in 1968).

When we calculate 95% confidence intervals for each year, the intervals have significant overlap. For independent samples, unless the overlap is small, this usually indicates that the difference will not be statistically significant in the corresponding t-test.

However, in the case of paired samples, such as in this data, we cannot base our conclusion on results which assume independent samples. We are interested specifically in the trend within each city. Overall, is there a change in the LFPR values?

Labor Force Participation Rate



Before conducting the paired t-test, let's look at a few additional types exploratory data analysis.

On the left we have the SPSS results for a histogram and normal QQ-plot of the LFPR values for 1968.

On the right we have the same results using SAS for 1972.

For both years, there seems to be one low outlier but in general the distributions are reasonably normally distributed.

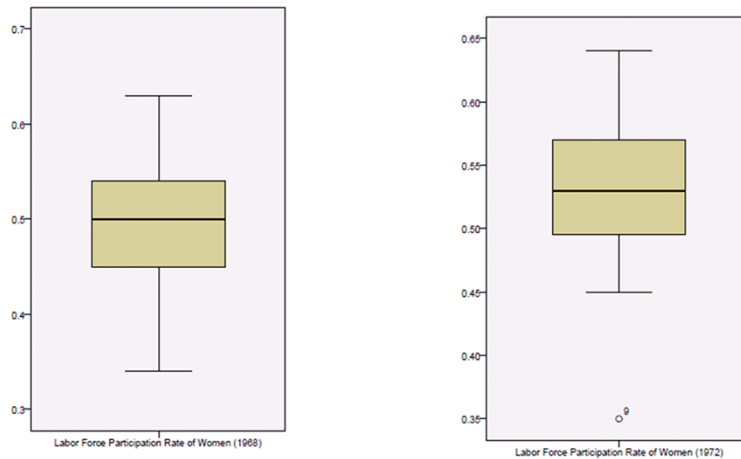
The histograms, give us a good overall picture of how the values of the variables are distributed for this sample of cities.

The normal QQ-plots are mostly used to investigate the validity of normality assumptions required by inferential methods you wish to use. For the moment, we aren't particularly concerned with the normality but we can see that both are approximately normal.

For the moment, it is difficult to make any comparisons based upon the histograms, especially when taken from two different packages.

In addition, in order to answer our question, we need to consider the pairing.

Labor Force Participation Rate



In the boxplots, from SPSS, we can see that the minimum in 1968 (which was Pittsburg with 0.34) is not considered an outlier whereas in 1978, the minimum is an outlier (this was St. Louis with 0.35)

LFPR: Data

City	LFPR72	LFPR68	Diff
N.Y.	0.45	0.42	0.03
L.A.	0.50	0.50	0
Chicago	0.52	0.52	0
Philadelphia	0.45	0.45	0
Detroit	0.46	0.43	0.03
San Francisco	0.55	0.55	0.00
Boston	0.60	0.45	0.15
Pitt.	0.49	<u>0.34</u>	0.15
St. Louis	<u>0.35</u>	0.45	-0.1
Connecticut	0.55	0.54	0.01

City	LFPR72	LFPR68	Diff
Wash., D.C.	0.52	0.42	0.10
Cinn.	0.53	0.51	0.02
Baltimore	0.57	0.49	0.08
Newark	0.53	0.54	-0.01
Minn/St. Paul	0.59	0.50	0.09
Buffalo	<u>0.64</u>	0.58	0.06
Houston	0.50	0.49	0.01
Patterson	0.57	0.56	0.01
Dallas	<u>0.64</u>	<u>0.63</u>	0.01

To begin our paired analysis, we can calculate the differences for each city.

You can see that most values are positive with a few negative and some with no measurable change.

Paired t-test

- $H_0: \mu_d = 0$
- $H_a: \mu_d \neq 0$
- Where μ_d = population mean of the difference in labor force participation rates among women for US cities between 1972 and 1968.

Mean	95% CL Mean	Std Dev	95% CL Std Dev
0.0337	0.00489	0.0625	0.0597

DF	t Value	Pr > t
18	2.46	0.0244

One-Sample Test

	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence ...	95% Confidence ...
					Lower	Upper
Difference in LFPR (1972 - 1968)	2.458	18	.024	.03368	.0049	.0625

Although researchers may be interested in showing an increase, in keeping with the most common analysis in practice we will conduct a two-sided test.

Our hypotheses will be that $\mu_{\text{sub-d}} = 0$ for the null hypothesis and $\mu_{\text{sub-d}} \neq 0$ for the alternative.

It is important to understand exactly what the symbolic parameters in your hypotheses mean in context.

In this case, we can define $\mu_{\text{sub-d}}$ to be the population mean of the difference in labor force participation rates among women for US cities between 1972 and 1968.

We have a small sample ($n = 19$) so we will need to investigate the normality assumption.

We will be using the sample standard deviation as we do not know the population standard deviation of the differences under study – thus we will be conducting a t-test.

In particular, we are conducting the paired t-test, however, this is the same process as a one-sample t-test except that usually the mean difference specified in the null hypothesis for a paired t-test is zero whereas for a one-sample t-test the null value is not usually zero.

Partial output for both packages is provided. We find

- A test statistic of $t = 2.46$ from SAS and 2.458 from SPSS
- A p-value of 0.0244 from SAS and 0.024 from SPSS
- The degrees of freedom are stated to be 18 indicating that n is 19 (as it should be)

Since the p-value is less than 0.05, we can reject the null hypothesis. We can say:

There is enough evidence to conclude that the population mean of the difference in labor force participation rates among women for US cities between 1972 and 1968 is not zero.

Although that interpretation is completely accurate, it may be re-worded for easier understanding as:

There was a statistically significant change in the population mean labor force participation rate among women for US cities between 1968 and 1972.

Now let's investigate the change:

- The estimated mean difference is 0.0337 from SAS and 0.03368 from SPSS
- The 95% confidence interval rounded to three decimal places is (0.005, 0.0625)

We can interpret this by saying:

Based upon our data, we estimate that the population mean labor force participation rate among women for US cities increased by 0.034 between 1968 and 1972. The 95% confidence interval suggests this value could be as low as 0.005 to as high as 0.0625.

Or we could simply say:

We are 95% confident that the population mean labor force participation rate among women for US cities increased by between 0.005 to 0.0625 from 1968 to 1972.

Remember that both our confidence intervals and hypothesis tests are about the population NOT our current sample – we know exactly what happened in our sample.

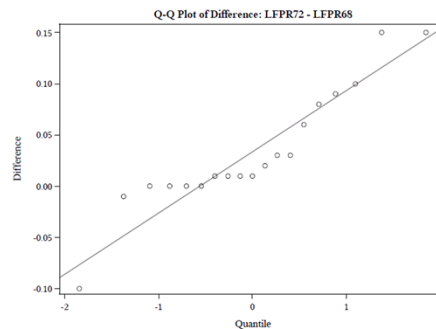
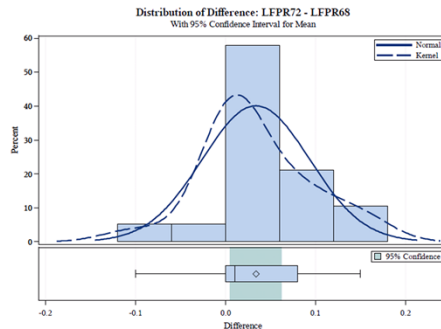
As always, it is possible that we have made an error.

For this hypothesis test, we could have claimed there was a change in the population mean labor force participation rate when in fact there was not which would be a Type I error.

And for the confidence interval, similarly, it is possible that it does not contain the true value.

We know that if we repeated this process, 95% of the time, the interval we obtain from this process would capture the target value but we cannot know if our current interval from 0.005 to 0.065 contains the true mean or not!

Labor Force Participation Rate



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between Labor Force Participation Rate of Women (1972) and Labor Force Participation Rate of Women (1968) equals 0.	Related-Samples Sign Test	.007 ¹	Reject the null hypothesis.
2	The median of differences between Labor Force Participation Rate of Women (1972) and Labor Force Participation Rate of Women (1968) equals 0.	Related-Samples Wilcoxon Signed Rank Test	.011	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

¹Exact significance is displayed for this test.

Tests for Location: Mu0=0				
Test	Statistic	p Value		
Student's t	t	2.457704	Pr > t	0.0244
Sign	M	5.5	Pr >= M	0.0074
Signed Rank	S	46	Pr >= S	0.0062

SAS does a better job than SPSS of providing details that help to validate the assumptions. In SPSS you would need to analyze the differences yourself where in SAS we obtain these graphs automatically upon conducting a paired t-test.

We can see that the distribution of the differences is somewhat not normal, however for such a small sample size, this would not be unexpected coming from a normal population.

It would be reasonable to apply the paired t-test.

If you are concerned, you could also apply the sign test and the Wilcoxon signed-rank test.

Both were applied in SAS and SPSS and both are statistically significant lending support to our conclusion of a statistically significant change.

You might notice that the p-value for the signed-rank test is different in SAS and SPSS. SAS uses the exact p-value where SPSS uses an asymptotic approach – which for small sample sizes may not be very accurate. There may be a way to find the exact p-value in SPSS but it wasn't a direct option that I could find.