

TWO VARIABLE STUDY PROPOSAL FORMAT

For this project, you can choose among these topics:

- Comparing population means, independent samples.
- Comparing population proportions, independent samples.
- Test for linear correlation, matched sample.

Here are some sample study goals (please think of something different for your study!)

- *Comparing means:* The purpose of this study is to detect a significant difference between the mean girth of an evergreen tree and the mean girth of a deciduous tree in Sacramento, with 10% level of significance.
- *Comparing proportions:* We will attempt to show that the proportion of 2-story houses in East Sacramento is greater than the proportion of 2-story houses in West Sacramento, with 5% level of significance.
- *Testing for correlation:* This is a proposal to test for a linear correlation between the length and the width of cars parked in Alhambra Triangle neighborhood, with 1% level of significance.

Your proposal should have the same 5 sections as the single sample project, and the only difference is that now you may have to talk about two populations and two samples, or else about two variables to measure.

Goal. State which population parameters you are trying to compare, or else which variables you are testing for correlation, and stipulate the desired confidence level α . State your null and alternative hypotheses.

Population(s), Sampling frame, Sampling method, Discussion. Same as for a single sample study, though possibly a different description or procedure for each population.

STATISTICAL STUDY PUBLICATION FORMAT

Amended proposal. Resubmit your original proposal, after amending it in response to the instructor's suggestions. Do not change any important parameters of your study, such as the intended sample size or the confidence level, without discussing it with the instructor. If you do, it may appear as an attempt to doctor the outcome of the study.

Data analysis. Provide full raw data in a separate spreadsheet file. State the relevant sample statistics. State the outcome of the test and the p -value of the test.

Discussion. This is your time to be both creative and reflective. Mention all the things that went wrong, but also pat yourself on the back for all the things that went right. Re-examine the assumptions and discuss how well they were met in your sample. For proportion studies, you would want large samples and a much larger populations, especially if the point estimates are extreme. For mean studies, you would like to see approximately normal populations or else large sample sizes. If you test for correlation, approximate normality is desired for both variables; also, you can discuss possible causation or lack thereof.

EXAMPLE: TEST FOR CORRELATION PROPOSAL

Goal. This is a proposal to test for a linear correlation between the length and the width of cars parked in Alhambra Triangle neighborhood, with 1% level of significance.

Population. Every 4-wheeled motorized vehicle parked on a public street in the Alhambra Triangle will be considered for this study.

Sampling frame. Since we cannot predict which cars will be parked in the testing area during the sample collecting stage, we will have to use a multi-stage approach, taking a cluster sample of streets in Alhambra triangle, and then taking a systematic sample of parked vehicles on each selected street.

Sampling method. First, we will make a list of all streets in Alhambra Triangle (see map image within the Example CI proposal), and choose a simple random sample of 8 random streets, using R for random number generation.

We will walk down each street once, down the odd-numbered side, during a convenient time, and measure every 6th car parked on that side of the street using the systematic sample technique. To pick the first car to be measured, a fair six-sided die will be rolled.

We will measure the dimensions of a car by laying a measuring tape on the ground about 1 foot away from the car. Two measurements will be taken: length and width.

The sample size cannot be known in advance, but we expect to sample at least 3 cars on most streets, giving us a good chance of $n > 20$.

Discussion.

- (a) If all selected streets are very short, our sample size may end up being too small for a meaningful result.
- (b) Bias is expected due to skipping all cars parked in yards and private driveways, though it is not clear what effect it might have on the correlation strength.
- (c) We expect to detect a significant positive correlation, since we expect larger cars to have greater width and length; the general size of a vehicle is hard to define, but may well be the lurking variable in this case.

EXAMPLE: TEST FOR CORRELATION PUBLICATION

Data. Alhambra Triangle street list:

R St, S St, Serra Way, T St, U St, 30th, 31st, 34th, 35th, Stockton Blvd

For our cluster sample, we picked all of the above streets except for 34th and Stockton Blvd.
The raw sample data can be found in a separate file: **sample-data-correlation.csv**

Data analysis.

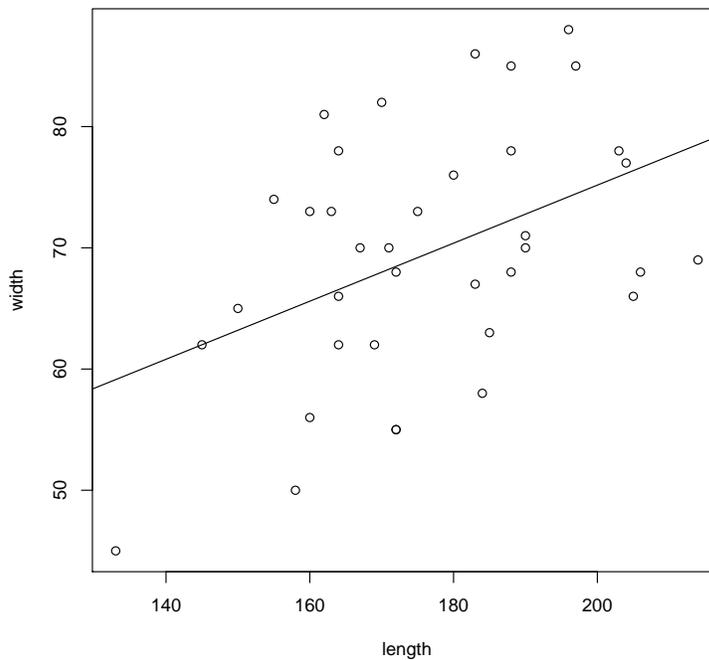
$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

This is a two-tail test with $\alpha = 0.01$

Sample size: $n = 37$

Pearson's Correlation Coefficient: $r = 0.4360384$



The sampling distribution is t with 35 degrees of freedom.

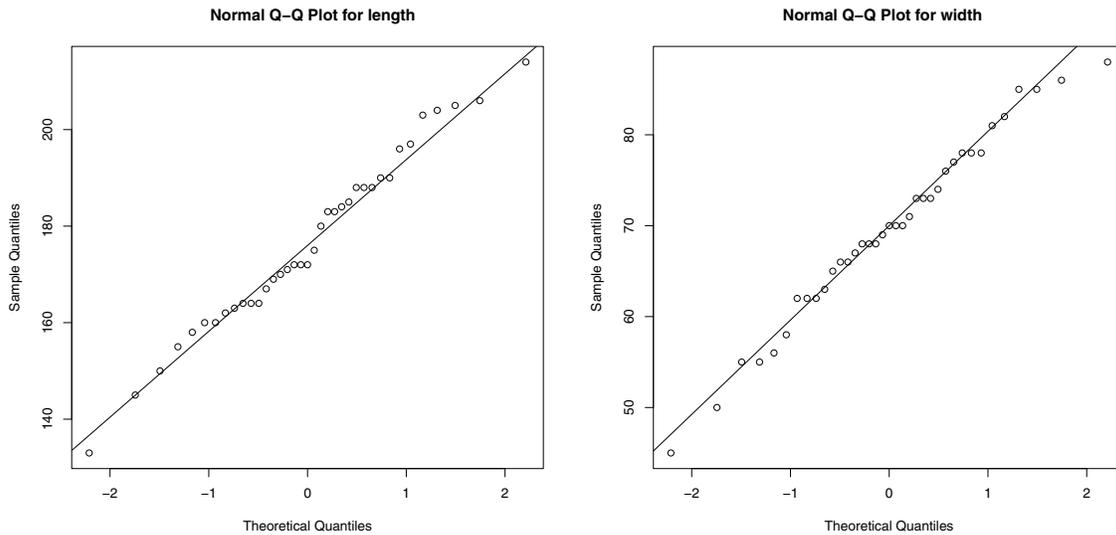
Test statistic: $t_0 = 2.866494$

Critical values of the sampling distribution: ± 2.723806

p -value is 0.006980393

Conclusion. We can reject H_0 and conclude that there is sufficient evidence to support the claim that the length and the width of a car are linearly correlated.

Discussion. The following QQ-plots summarize the strength of normality in the sample data:



- (a) Though the sample size is relatively small, the confidence in our results is improved by the approximately normal distribution of the measured variables, seen in the plots above.
- (b) Just as we presumed before commencing the study, there seems to be a positive association. A future study may well take a larger sample and seek the evidence of a strictly positive trend.