Lab One, Part One

Samantha Williams, Varune Maharaj, Wilford Bradford, Dmitri Zadvornov

Feb 26, 2022

Contents

1 Part 1: Foundational Exercises

1.1 Professional Magic

Utilizing available information, let $U_i = X_i + Y_i$. Then with a given joint distribution of X_i and Y_i we can calculate distribution function of U_i as follows:

$P[U = 0] = P[X_i = 0, Y_i = 0] = \frac{p}{2}$

$P[U = 1] = P[X_i = 1, Y_i = 0] + P[X_i = 0, Y_i = 1] = \frac{1-p}{2} + \frac{1-p}{2} = 1 - p$

$P[U = 2] = P[X_i = 1, Y_i = 1] = \frac{p}{2}$

Then the distribution of U_i is:

$$f_U(u) = \begin{cases} \frac{p}{2}, u = 0 \\ 1-p, u = 1 \\ \frac{p}{2}, u = 2 \\ 0, otherwise \end{cases}$$

Finally, $X_1 + Y_1 + X_2 + Y_2 + X_3 + Y_3 = U_1 + U_2 + U_3$.

1. Calculating type 1 error rate - α

$\alpha = P[\text{rejecting } H_0 | H_0] = P\left[\sum_{i=1}^{3} U_i = 0 \cup \sum_{i=1}^{3} U_i = 6 \,\middle|\, p = \frac{1}{2}\right]$.

Given that events $\sum_{i=1}^{3} U_i = 0$ and $\sum_{i=1}^{3} U_i = 6$ are mutually exclusive:

$P\left[\left(\sum_{i=1}^{3} U_i = 0\right) \cup \left(\sum_{i=1}^{3} U_i = 6\right) \,\middle|\, p = \frac{1}{2}\right] =$
$P\left[\sum_{i=1}^{3} U_i = 0 \,\middle|\, p = \frac{1}{2}\right] + P\left[\sum_{i=1}^{3} U_i = 6 \,\middle|\, p = \frac{1}{2}\right] =$

$2\left(\prod_{i=1}^{3}\frac{p}{2}\right)\Big|_{p=\frac{1}{2}} = 2\frac{p^3}{8}\Big|_{p=\frac{1}{2}} = \frac{1}{4}\frac{1}{8} = \frac{1}{32} = 0.03125$

2. Calculating power $(1-\beta)$ of the test

$power = P[\text{rejecting } H_0|H_a] = P\left[\sum_{i=1}^{3}U_i = 0\,\Big|\, p=\frac{3}{4}\right] + P\left[\sum_{i=1}^{3}U_i = 6\,\Big|\, p=\frac{3}{4}\right] =$
$2\left(\prod_{i=1}^{3}\frac{p}{2}\right)\Big|_{p=\frac{3}{4}} = 2\frac{p^3}{8}\Big|_{p=\frac{3}{4}} = \frac{1}{4}\frac{27}{64} = \frac{27}{256} = 0.1054$

1.2 Wrong Test, Right Data

Imagine that your organization surveys a set of customers to see how much they like your regular website, and how much they like your mobile website. Suppose that both of these preference statements are measured on 5-point Likert scales.

A Likert scale is one where a person is provided ordered categories that range from lowest to highest. You can read more about them in this seminal research design text by Fowler, or this brief overview. If you were to run a paired t-test using this data, what consequences would the violation of the metric scale assumption have for your interpretation of the test results? What would you propose to do to remedy this problem?

1.2.1 General requirements of a paired t-test:

- Must be IID
- Data must be metric
- Each paired measurement must be obtained from the same customer
- The distribution of differences between the paired measurements are sufficiently normally distributed (more important for small sample size than for larger sample size)

1.2.2 Consequences of a paired t-test using the data provided:

A paired t-test is used to test if the *mean* difference between two pairs of measurements is zero or not. This is a statistical test that is not designed for this type of data.

We are provided ordinal data that ranks a respondent's opinion of a company's mobile and regular website on 5-point Likert scale. It is important to mention that the ordinal scale does not capture the true distance between consecutive levels on the scale, in which case the mean calculation and mean comparison do not have a meaning, potentially misrepresenting both quantities in either direction - understating or overstating, as well as misrepresenting the center and the spread of the distribution.

Performing a test that requires metric data with an ordinal-scale data would result in an inaccurate description of the relationship between customer preference of the company's mobile site vs. the regular website. Because we cannot accurately measure the difference in scores, we may assume that the data is not a normal or t- distribution regardless of the sample size.

1.2.3 Solution to properly asses this data set:

Perform a **Wilcoxon sign test** which compares the sample *median* against a hypothetical median where the null hypothesis for this test is the medians of two samples are equal. This test is non-parametric and requires ordered categorical variables that do not have to be metric (ordinal data) that is IID. Given the information provided this would be a better statistical test to explore, whether the sites were not equally liked in one or both directions (mobile website vs. regular website) by the customers sampled.

1.3 Test Assumptions

1.3.1 World Happiness

Given that the two-sample t-test is utilized, the following assumptions must hold:

1. Underlying random variables X and Y are metric
2. X_i and Y_i are IID
3. Size of the sample is large enough relative to the distribution degree of skewness
4. Equal Variance for underlying group distributions

1. Metric Scale: Reviewing underlying data, the variable for happiness is measured on the ordinal scale, as it represents the level of happiness on the scale of 0-10, where the distance of 1 between levels 9 and 10 would not necessarily be the same as the distance of 1 between 0 and 1. In fact, negative feelings are affected by negativity bias, which amplifies the magnitude of negative feelings. Further, the t-test statistic is calculated based on the difference of means, which in itself requires at least interval scale. The type of data provided violates this assumption.

2. Sample value pairs are IID: Gallup surveys are based on randomly selected and nationally representative samples from more than 140 countries. We can technically assume that they are random enough, but it is also important to note that not all countries are distinct enough based on culture, historical diffusion of language, culture, and customs of neighboring countries, as well as effects from geopolitical structures, such as trade unions and other.

3. Size of the sample is large relative to degree of skewness of the distribution:

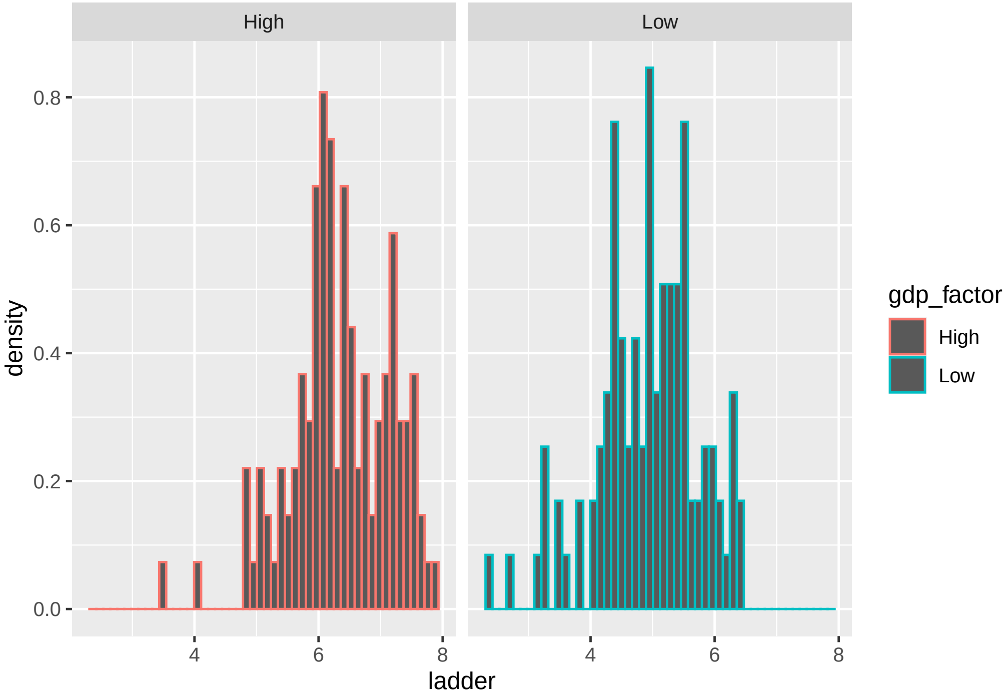
We can observe that the lowest category count is 105, which is generally a good sample size for distributions that are not highly skewed.

	x
High	121
Low	105
NA's	13

The skewness values for underlying distributions are not extreme and indicate slight left skew of the distributions.

x
-0.5005567
-0.5395703

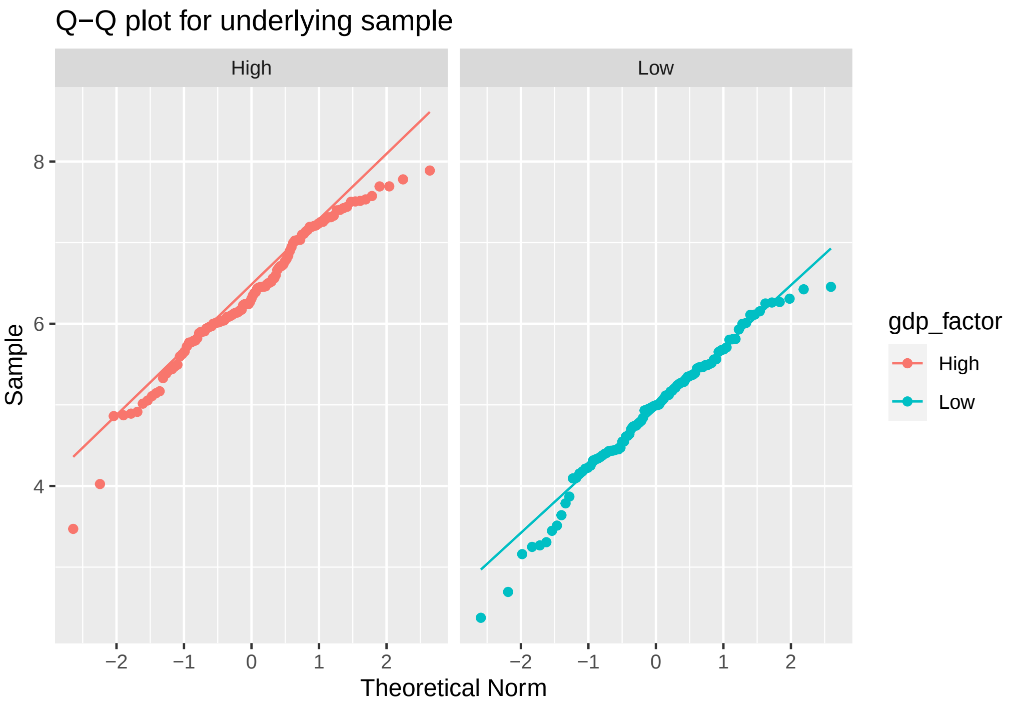
Looking at the distribution shapes based on the density histogram, we can confirm that both samples are slightly skewed to the left.



4. **Equal Variance for underlying group distributions:** Taking a glance at the plots above we can also conclude that the spread (variance) for both is quite similar. Also calculating sample variance we can confirm that it is very similar between samples. This assumption is satisfied.

x
0.6438919
0.6655736

Finally, comparing underlying distributions to the normal we see that skewness have limited effects with left tail being slightly heavier, while the right tail is slightly lighter than the normal distribution.



For the left tail:

So y or sample has smaller lower quantiles than normal - suggesting heavier left tail (e.g. more data on the left than for normal tail), but not too extreme visually.

For the right tail:

The sample has smaller quantiles than normal - suggesting lighter right tail than for normal distribution (e.g. more data based on quantiles on the right for normal than sample), but not too extreme visually.

Under CLT assumptions and good sample size we should be OK for this assumption.

1.3.2 Legislators

1.3.2.1 Assumptions for a Wilcoxon rank-sum test:

1. Data must be at least ordinal
2. Sample pairs are IID
3. No substantial differences in group sample sizes
4. Continuous data

A quick review of the data shows us that there are 3 levels for the party. Birth information is represented by the birthday. We can operationalize the degree of being old by either using POSIX or calculating age to a specific fixed date. The latter is a cleaner approach.

```
## data132
##
##  3  Variables      536  Observations
## --------------------------------------------------------------------------------
## party
##        n  missing distinct
##      536        0        2
##
## Value        Democrat Republican
## Frequency         272        264
## Proportion      0.507      0.493
## ---------------------------------------------------------------------------
## birthday
##          n    missing   distinct       Info       Mean        Gmd        .05
##        536          0        528          1 1961-12-24       4958 1943-09-10
##        .10        .25        .50        .75        .90        .95
## 1947-05-01 1953-04-18 1961-04-03 1970-09-09 1978-10-26 1982-08-26
##
## lowest : 1933-06-09 1933-06-22 1933-09-17 1934-05-06 1934-11-17
## highest: 1988-02-17 1988-03-12 1989-02-01 1989-10-13 1995-08-01
## ---------------------------------------------------------------------------
## age
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      536        0      509        1    60.22    13.58    39.53    43.36
##      .25      .50      .75      .90      .95
##    51.50    60.95    68.91    74.88    78.52
##
## lowest : 26.59 32.40 33.09 33.99 34.05, highest: 87.34 87.87 88.51 88.74 88.78
## ---------------------------------------------------------------------------
```

party	birthday	age
Democrat	1952-11-09	69.35
Democrat	1958-10-13	63.42
Democrat	1943-10-05	78.45
Democrat	1947-01-23	75.15
Democrat	1960-04-13	61.92
Democrat	1933-06-22	88.74

1. Data must be at least ordinal: Age data is metric, as intervals between dates carry appropriate meaning, zero has a meaning of not starting life, and the idea of someone 10 years older has a meaning of having 10 years more to the life experience than the other person. As the data scale is above ordinal, the assumption is satisfied.

2. Sample pairs are IID: legislators-current data set holds current population information for current serving members of the congress. As such, we are working with the population, and not sample distribution. Independence is required for sample to provide maximum information about the underlying population, but in this case we already have all of the members of the population. As such, we can use the population as is and compare population parameter between groups directly, or draw random samples from existing population set to directly satisfy this requirement.

3. No substantial differences in group sample sizes: Based on the summary of the data we obtained before, we have group sizes that do not deviate substantially - 272 vs. 264. This assumption is satisfied.

4. Continuous data: The age data is continuous, so this specific assumption is satisfied.

1.3.3 Wine and health

The wine data set contains observations of variables related to wine consumption for 21 countries. You would like to use this data to test whether countries have more deaths from heart disease or from liver disease.

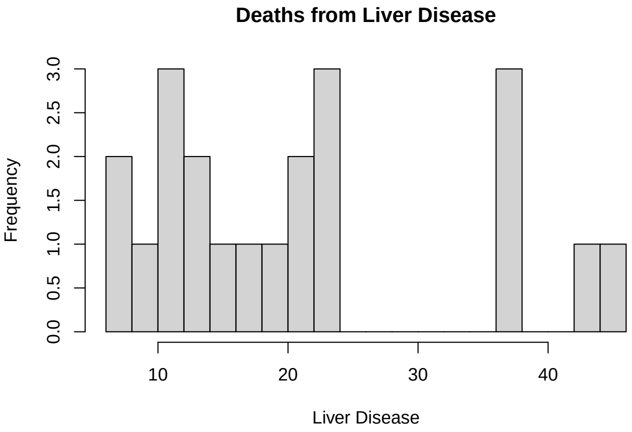
List all assumptions for a signed-rank test. Then evaluate each assumption, presenting evidence based on your background knowledge, visualizations, and numerical summaries.

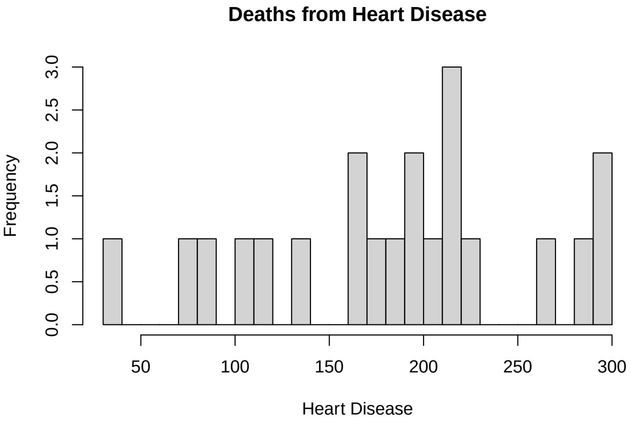
1.3.3.1 Singed-Rank Test Assumptions:

country	alcohol	deaths	heart	liver
Australia	2.5	785	211	15.3
Austria	3.9	863	167	45.6
Belg/Lux	2.9	883	131	20.7
Canada	2.4	793	191	16.4
Denmark	2.9	971	220	23.9
Finland	0.8	970	297	19.0

country	alcohol	deaths	heart	liver
Length:21	Min. :0.600	Min. : 680	Min. : 36.0	Min. : 6.50
Class :character	1st Qu.:1.200	1st Qu.: 751	1st Qu.:131.0	1st Qu.:11.20
Mode :character	Median :1.900	Median : 806	Median :191.0	Median :19.00
NA	Mean :2.838	Mean : 830	Mean :183.3	Mean :21.03
NA	3rd Qu.:2.900	3rd Qu.: 916	3rd Qu.:220.0	3rd Qu.:23.90
NA	Max. :9.100	Max. :1000	Max. :300.0	Max. :45.60

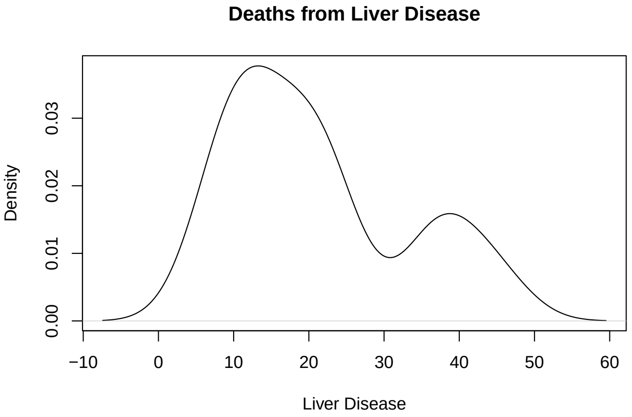
Non-Parametric: The data is non-parametric with a small sample size (21) that does not have a normal distribution.

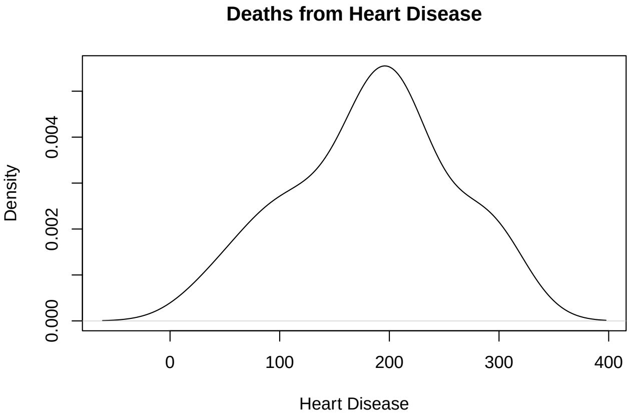

Deaths from Liver Disease


Deaths from Heart Disease

Paired: The wine data is unpaired. There is no dependency between the people who die from liver disease versus heart disease.

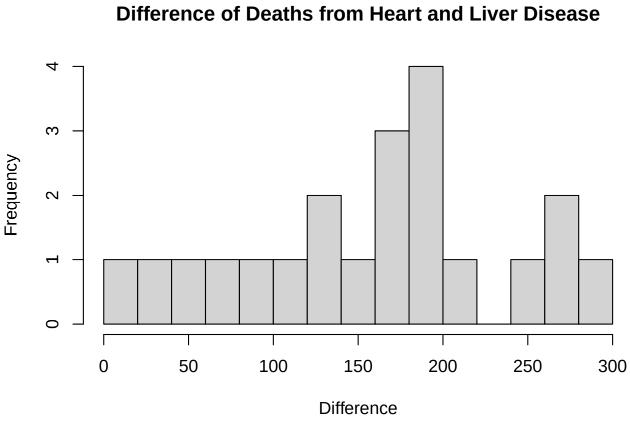
Must be IID: This data set may not be IID. While the wine data may be independent, it is not identically distributed across liver and heart disease.

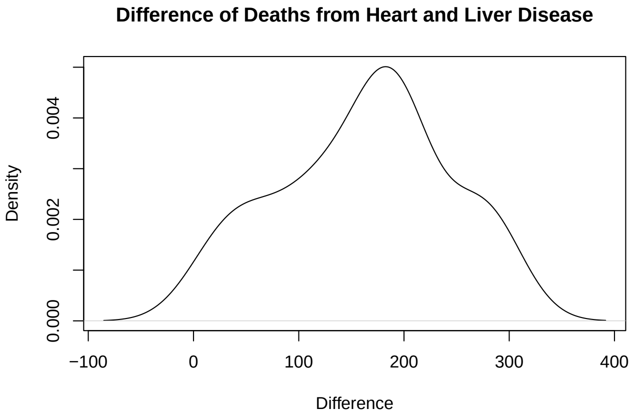

Deaths from Liver Disease


Deaths from Heart Disease

Data must be metric: The wine data set is metric with counts of the number of deaths from liver, heart and general deaths per 100,000 people. Alcohol is counted in liters of wine, per capita.

Difference is symmetric: The wine data is not symmetric based on the plot of the distribution of the difference of heart and liver deaths.


Difference of Deaths from Heart and Liver Disease


Difference of Deaths from Heart and Liver Disease

1.3.4 Attitudes toward the religious

1.3.4.1 Assumptions for a paired t-test:

1. Metric scale data
2. Paired - dependence within samples
3. Must be IID
4. Data must be metric
5. Difference of between sample values must be sufficiently normal

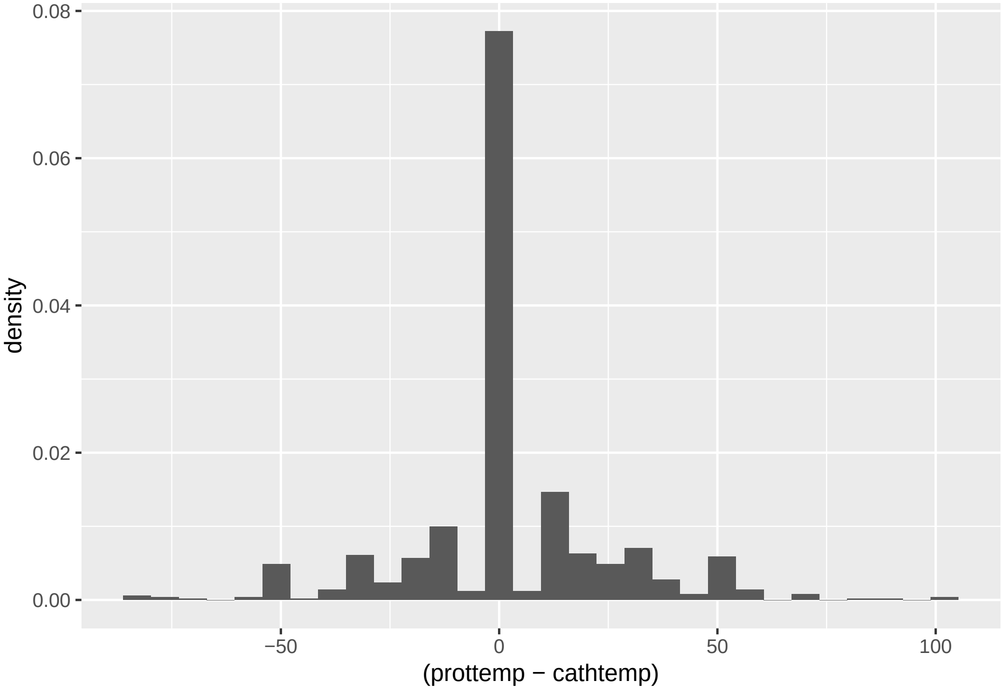
1. Metric scale data: The setup for rating of feelings by the respondents is using thermometer scale, which is on the interval scale. It is not of ratio scale with respect to zero, so the ratios and division may not have a strong meaning.

```
## data.1.3.4
##
##  5  Variables      802  Observations
## --------------------------------------------------------------------------------
## ...1
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      802        0      802        1    401.5    267.7    41.05    81.10
##      .25      .50      .75      .90      .95
##   201.25   401.50   601.75   721.90   761.95
##
## lowest :   1   2   3   4   5, highest: 798 799 800 801 802
## --------------------------------------------------------------------------------
## year
##        n  missing distinct     Info     Mean      Gmd
##      802        0        1        0     2004        0
##
## Value      2004
## Frequency   802
## Proportion    1
## --------------------------------------------------------------------------------
## id
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      802        0      802        1     1382    903.8    150.2    295.3
##      .25      .50      .75      .90      .95
##    728.8   1373.5   2053.5   2462.4   2621.8
##
## lowest :    4    7    9   14   21, highest: 2800 2803 2805 2806 2808
## ---------------------------------------------------------------------------------
## prottemp
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      802        0       25    0.931    65.56    23.74       40       50
##      .25      .50      .75      .90      .95
##       50       60       85      100      100
##
## lowest :   0   1   2   5   7, highest:  85  90  95  99 100
## ---------------------------------------------------------------------------------
## cathtemp
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      802        0       27    0.962    63.16    25.74       25       40
##      .25      .50      .75      .90      .95
##       50       60       85      100      100
##
## lowest :   0   1   2   5   7, highest:  85  90  95  99 100
## ---------------------------------------------------------------------------------
```

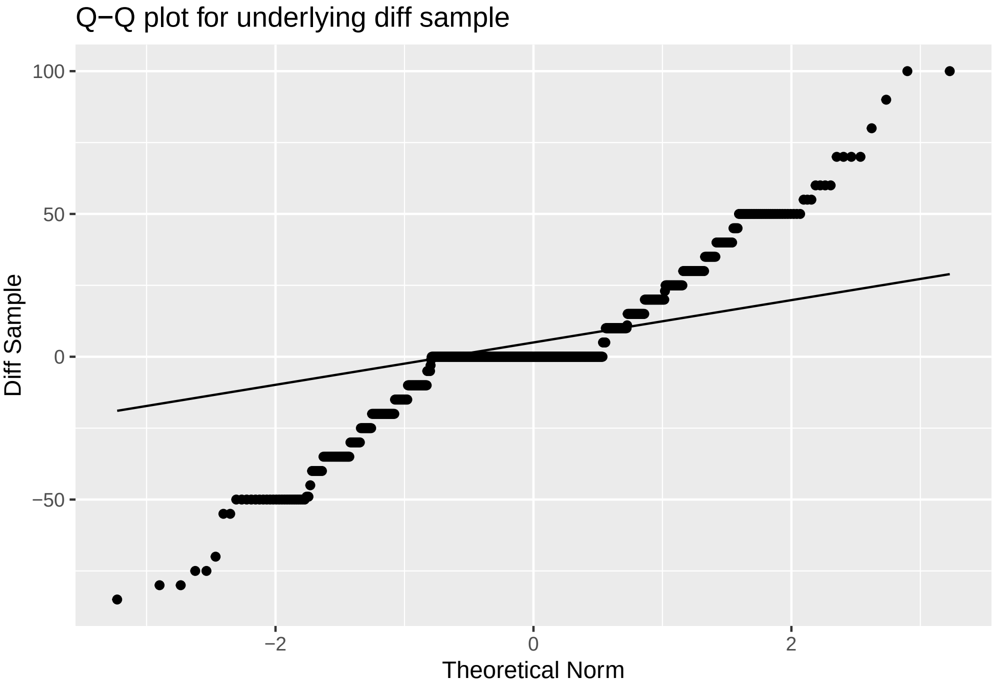
2. Paired - dependence within samples: As response is given by the same subject to 2 types of stimuly, the dependence assumption is satisfied.

3. Must be IID: Survey approach uses area probability design that randomly selects respondents in households across US. This approach should provide sufficiently independent sample.

4. Difference of between sample values must be sufficiently normal: From the density histogram we can see potentially heavy tails.



The Q-Q plot confirms significant deviations from normal distribution in terms of weight of tails. This would mean that the CLT convergence would be much slower and require much large sample size. At the given sample size this assumption does not hold.


Q−Q plot for underlying diff sample
Diff Sample
100
50
0
−50
−2
0
2
Theoretical Norm