Basic Course on R: Hypothesis Testing and Confidence Intervals 1

Elizabeth Ribble*

28 Oct - 1 Nov 2019

Contents

*emcclel3@msudenver.edu

1 Summary Statistics

```
Puromycin
```

```
##    conc rate     state
## 1  0.02   76   treated
## 2  0.02   47   treated
## 3  0.06   97   treated
## 4  0.06  107   treated
## 5  0.11  123   treated
## 6  0.11  139   treated
## 7  0.22  159   treated
## 8  0.22  152   treated
## 9  0.56  191   treated
## 10 0.56  201   treated
## 11 1.10  207   treated
## 12 1.10  200   treated
## 13 0.02   67 untreated
## 14 0.02   51 untreated
## 15 0.06   84 untreated
## 16 0.06   86 untreated
## 17 0.11   98 untreated
## 18 0.11  115 untreated
## 19 0.22  131 untreated
## 20 0.22  124 untreated
## 21 0.56  144 untreated
## 22 0.56  158 untreated
## 23 1.10  160 untreated
```

```
attach(Puromycin)
```

1.1 Continuous data

Measures of center:

```
mean(rate)
```

```
## [1] 126.8261
```

```r
mean(rate, trim = 1/10)

## [1] 126.8947

median(rate)

## [1] 124
```

where the median is the middle of n numbers and the mean is defined as

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

Measures of spread:

```r
var(rate)

## [1] 2257.514

sqrt(var(rate))

## [1] 47.5133

sd(rate)

## [1] 47.5133
```

Note that the formula for standard deviation is

$$\mathrm{s}_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

(variance is s_x^2) so we could also make our own function:

```r
stdev <- function(x){ sqrt(sum((x-mean(x))^2)/(length(x)-1)) }
stdev(rate)

## [1] 47.5133

IQR(rate)

## [1] 67
```

```r
iqr <- function(x){
  q <- quantile(x)
  q[4]-q[2]
}
iqr(rate)

## 75%
##  67

abs(-1)

## [1] 1

mad(rate)

## [1] 51.891

MAD <- function(x){ median(abs(x-median(x)))*1.4826 }
MAD(rate)

## [1] 51.891
```

The value 1.4826 is a normalizing constant; it makes the MAD comparable to the standard deviation of a normal distribution.

Other summary statistics and useful functions:

```r
summary(rate)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    47.0    91.5   124.0   126.8   158.5   207.0

quantile(rate)

##    0%   25%   50%   75%  100%
##  47.0  91.5 124.0 158.5 207.0

quantile(rate, seq(0, 1, .1))

##    0%   10%   20%   30%   40%   50%   60%   70%   80%   90%  100%
##  47.0  68.8  84.8  97.6 113.4 124.0 140.0 154.4 159.6 198.2 207.0
```

```r
min(rate)
```

```
## [1] 47
```

```r
max(rate)
```

```
## [1] 207
```

```r
sum(rate)
```

```
## [1] 2917
```

```r
log(rate[1])
```

```
## [1] 4.330733
```

```r
log(rate[1], 10)
```

```
## [1] 1.880814
```

```r
log10(rate[1])
```

```
## [1] 1.880814
```

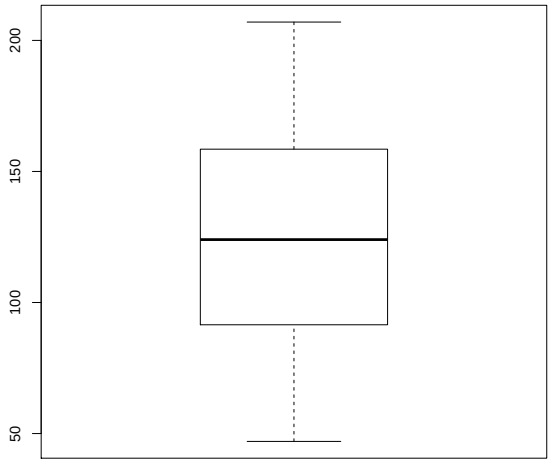
```r
sort(round(rate, -1))
```

```
##  [1]  50  50  70  80  80  90 100 100 110 120 120 120 130 140 140 150 160
## [18] 160 160 190 200 200 210
```

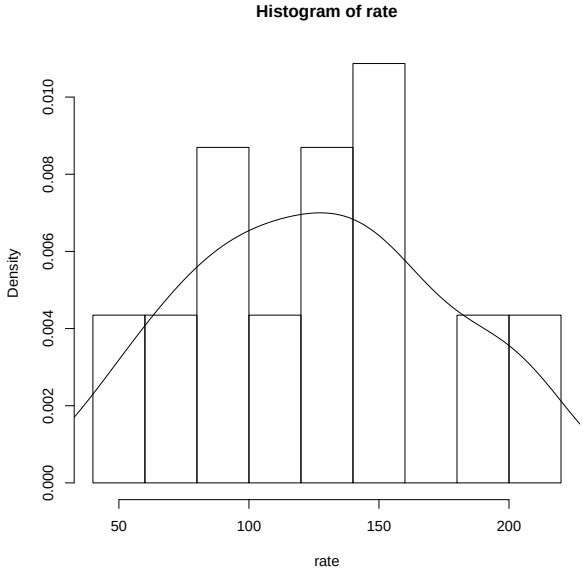
```r
stem(rate)
```

```
##
##   The decimal point is 2 digit(s) to the right of the |
##
##   0 |
##   0 | 557889
##   1 | 001222344
##   1 | 56669
##   2 | 001
```

```r
boxplot(rate)
```




```r
hist(rate, freq = F)
lines(density(rate))
```



Generating random variables and selecting random samples:

```
set.seed(123)
x <- rnorm(100)
stem(x)

##
##   The decimal point is at the |
##
##   -2 | 30
##   -1 | 75
##   -1 | 3321111000
##   -0 | 77777666665555
##   -0 | 44443333322222211100
##    0 | 01111122223334444444
##    0 | 55566677888999
##    1 | 0001122344
##    1 | 55678
##    2 | 122

y <- runif(100, 0, 1)
stem(y)

##
##   The decimal point is 1 digit(s) to the left of the |
##
##   0 | 12555666777
##   1 | 0004567777
##   2 | 022234566789
##   3 | 014556
##   4 | 0000000137788
##   5 | 022345689
##   6 | 023567779
##   7 | 0023345679
##   8 | 0022377888
##   9 | 2235667778

sample(rate)

##  [1] 131  76  98  84 200 191  97 158 159 124 139  67 152  86 107  51 160
## [18]  47 207 115 144 201 123
```

```r
sample(rate, size = 5)

## [1]  98  97  76 124 144

sample(rate, size = 5, replace = TRUE)

## [1] 152 158  84 159  98

sample(rate, size = 5, replace = TRUE,
      prob = runif(length(rate),0,1))

## [1] 201 107  76 152  51
```

1.2 Categorical data

```r
table(state)

## state
##   treated untreated
##        12        11

table(state)/length(state)

## state
##   treated untreated
## 0.5217391 0.4782609
```

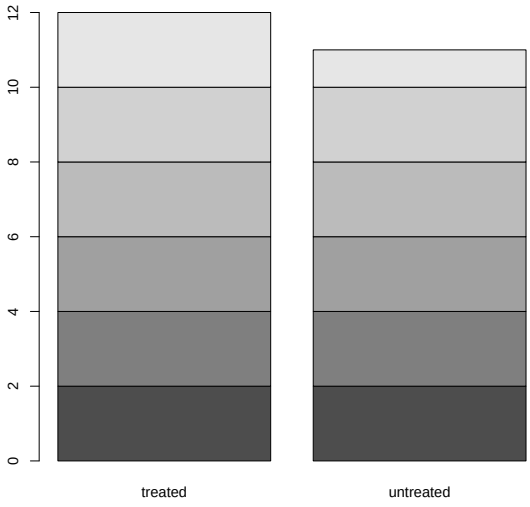
You can also make continuous data categorical:

```r
cats <- cut(rate, breaks = c(0,53,105,153,210))
table(cats)

## cats
##    (0,53]  (53,105] (105,153] (153,210]
##         2         6         8         7

cats

##  [1] (53,105]  (0,53]    (53,105]  (105,153] (105,153] (105,153] (153,210]
##  [8] (105,153] (153,210] (153,210] (153,210] (153,210] (53,105]  (0,53]
## [15] (53,105]  (53,105]  (53,105]  (105,153] (105,153] (105,153] (105,153]
## [22] (153,210] (153,210]
## Levels: (0,53] (53,105] (105,153] (153,210]
```

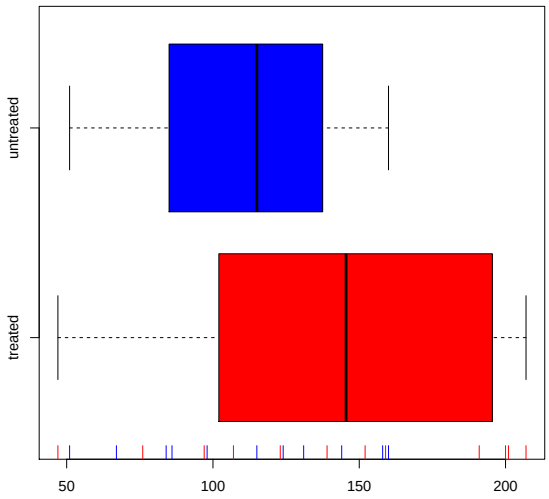
```
cats <- cut(rate, breaks = quantile(rate))
table(cats)

## cats
##  (47,91.5] (91.5,124]  (124,158]  (158,207]
##          5          6          5          6
```

1.3 Bivariate data

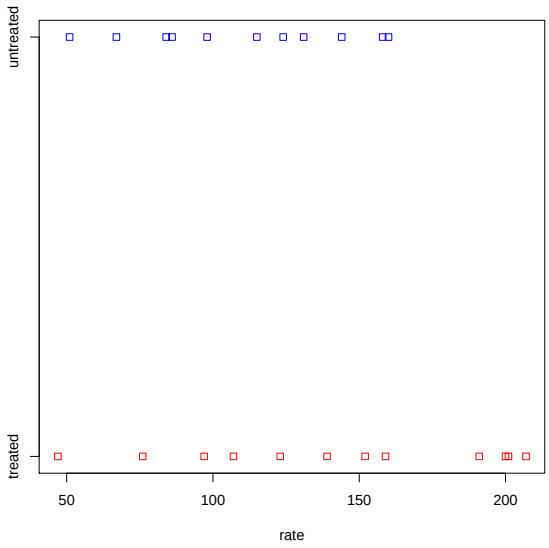
```
table(conc, state)

##       state
## conc   treated untreated
##   0.02       2         2
##   0.06       2         2
##   0.11       2         2
##   0.22       2         2
##   0.56       2         2
##   1.1        2         1

barplot(table(conc, state))
```



```
boxplot(rate~state, horizontal = TRUE, col = c("red", "blue"))
rug(rate[state=="treated"], col = "red")
rug(rate[state=="untreated"], col = "blue")
```




```
stripchart(rate~state, col = c("red", "blue"))
```



2 Hypothesis Testing and Confidence Intervals

Comparison is the most common basic principle in medical research. A statement about the truth is compared against a reference statement (the null).

- H_0: Null hypothesis, e.g. cholesterol is comparable between men and women
- H_1: Alternative hypothesis, e.g. men and women differ on average in cholesterol

The p-value is the probability of obtaining the observed data in the sample (or something more extreme than the observed data), given that the null hypothesis is true. The p-value is calculated based on a point estimate (e.g. mean) of the sample. The decision to reject a null hypothesis based on the p-value depends on a chosen α level.

	reality	
decision	H_0 is true	H_0 is false
do not reject H_0	yay!	type II error (β)
reject H_0	type I error (α)	yay!

For a given dataset and corresponding test statistic, we say the results of the test are "statistically significant" if the p-value is less than the pre-selected and fixed significance level, α.

Note that there is a distinction between significance and clinical relevance:

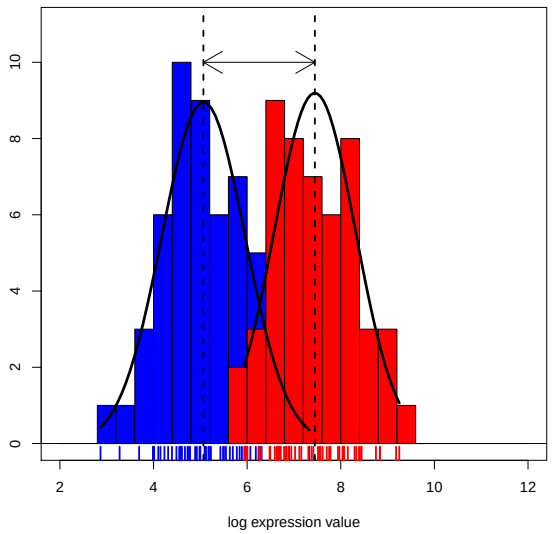
- If the sample size is large enough, even a small difference of 0.1 mmHg blood pressure can be *statistically significant* between groups, though it is not *relevant* from a clinical point of view.
- If the sample size is too small, even a sample mean of 150 mmHg can be *not statistically significantly different* from 130 mmHg, though 20 mmHg is *clinically relevant*.

A confidence interval (CI) is another way to show the reliability of a point estimate. The decision to reject or not reject the null hypothesis aligns with whether or not the CI contains the null value (e.g. H_0: mean = 0; if CI does not contain 0 then reject, otherwise do not reject H_0). In other words, the decision made by comparing the test statistic's p-value to α will be the same as a decision made using a $(1-\alpha)*100\%$ CI.

An interpretation of e.g. a 95% CI is "if the testing procedure were repeated on many (k) samples, the confidence intervals would encompass the true population parameter in 95% of the k samples" or, more abstractly, "we are 95% confident that the true [e.g. mean] lies in the confidence interval".

2.1 t-Test

The t-test is a statistical procedure used to test for the difference in means between two independent populations. The samples should come from normal distributions (can check using e.g. `qqnorm()`) and the variances from each population are assumed to be equal.



The t-test in R makes a correction that does not require equal variances of the two populations. The null hypothesis is that the difference between means is 0; the alternative is that this difference is not 0.

The t test statistic is defined as

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

and for large values of this test statistic we reject the null hypothesis that the difference in means is 0. A large value of this t will produce a small p-value which, again, is evidence against the null hypothesis.

The Welch-Satterthwaite correction for unequal variances is

$$\text{df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}$$

The two-sample α-level confidence interval is:

$$\bar{x_1} - \bar{x_2} \pm t_{\alpha/2,df}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where $t_{\alpha/2,df}$ is determined by the Student's-t distribution.

Here is an example of how to do this test in R. In the plotting lecture, it appeared that the OJ recipients in the sample had a larger tooth growth than the VC recipients, on average. We can test if this is true in the larger popluation using a t-test. Let's set H_0 to be that the two mean tooth lengths are the same and H_a will be that the two mean tooth lengths are not equal. Let's use a 0.10 significance level, which means the probability we falsely reject the null is 0.10. Note that the function `t.test` also provides a confidence interval - the confidence level specified should be $1 - \alpha$ for the interpretations to align.

```
ttooth <- t.test(ToothGrowth$len~ToothGrowth$supp, conf.level = 0.90)
ttooth

##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len by ToothGrowth$supp
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##  0.4682687 6.9317313
## sample estimates:
## mean in group OJ mean in group VC
##         20.66333         16.96333
```

Here the p-value is less than our chosen significance level so we reject the null. Therefore, at the 10% significance level, the data provide sufficient evidence that our alternative hypothesis, that there is a difference between using OJ and VC to grow teeth, is true.

Now consider the following example. Suppose we believe the mean reaction velocity in an enzymatic reaction will be higher in cells treated with Puromycin compared to cells not treated with Puromycin. Let's use the 0.05 significance level, allowing for a slightly smaller Type I error probability than in the previous example.

```
t.test(rate~state, alternative = "greater")

##
##  Welch Two Sample t-test
##
## data:  rate by state
## t = 1.6375, df = 19.578, p-value = 0.05875
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -1.677431       Inf
## sample estimates:
##   mean in group treated mean in group untreated
##                141.5833                110.7273
```

Here our p-value is greater than α so we do not reject the null hypothesis. We therefore conclude that, at the 5% significance level, the data do not provide sufficient evidence that Puromycin increases the mean reaction velocity.

We can extract statistics and calculate e.g. fold changes from the output:

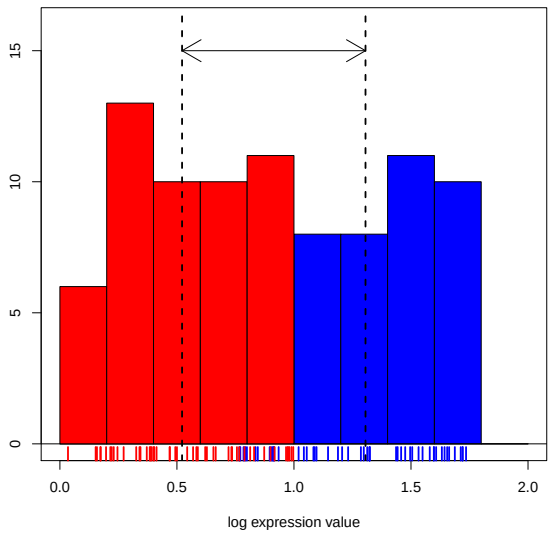
```
mytest <- t.test(rate~state, alternative="greater")
mytest$p.value

## [1] 0.05874922

mytest$conf.int

## [1] -1.677431       Inf
## attr(,"conf.level")
## [1] 0.95

mytest$estimate[2]/mytest$estimate[1]

## mean in group untreated
##               0.7820643
```

If samples are matched or paired (e.g. before/after), use argument `paired=TRUE`.

2.2 Mann-Whitney U Test

The Mann-Whitney U test (also known as the Wilcoxon Rank Sum test) is a non-parametric test for a location shift between two independent populations without assuming normality. However, the values should be sampled from two populations with

very similar distributions. The null hypothesis is that there is no shift in the centers of the distributions of the two populations; the alternative is that there is a shift.




```
wtest <- wilcox.test(rate~state,conf.int=TRUE)
wtest

##
##  Wilcoxon rank sum test
##
## data:  rate by state
## W = 88, p-value = 0.1896
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -10  75
## sample estimates:
## difference in location
##                     36

wtest$p.value

## [1] 0.1895867
```

The test statistic `W` in R is defined to be the sum of the ranks of one of the groups minus $n_1(n_1+1)/2$. The procedure to obtain a confidence interval is quite involved, but thankfully R will provide it to us upon request.

This bit of code shows how to calculate the W test statistic:

```
n1 <- length(state[state=="treated"])
n1

## [1] 12

rank(rate)

##  [1]  4  1  7  9 11 14 18 16 20 22 23 21  3  2  5  6  8 10 13 12 15 17 19

rank(rate)[state=="treated"]

##  [1]  4  1  7  9 11 14 18 16 20 22 23 21

sum(rank(rate)[state=="treated"])-n1*(n1+1)/2

## [1] 88
```