Significance Tests

Ira Sharenow

January 7, 2019

R Markdown

This paper shows how to perform some basic statistical tests using the R statistical programming language.

1. One sample tests
    1. t-test + graph (Example 1)
    2. Wilcoxon test (Example 2)
2. Two sample tests
    1. t-test + graph (Example 3)
    2. paired t-test (Example 4)
    3. binomial test (Example 5)
    4. prop test (Example 6)
3. chi-squared test of independence (Example 7)
4. One-way ANOVA (Example 8)
5. Two-way ANOVA (Example 9)
6. Power tests
    1. power.t.test (Example 10)

```
# One Sample Tests

# Example 1

# A call center has been getting about 100 calls per day during weekdays.
# Syed claims he has a marketing
# strategy that will produce more calls. His company gives it a try for two weeks (10 days).
# Below is the data. Was Syed's method effective?



callsS = c(112, 99, 105, 103, 108, 95, 116, 97, 106, 108, 94)
df1 = data.frame(callsS = callsS)

# perform a one-sided t-test

mean(callsS)
```

```
## [1] 103.9091
```

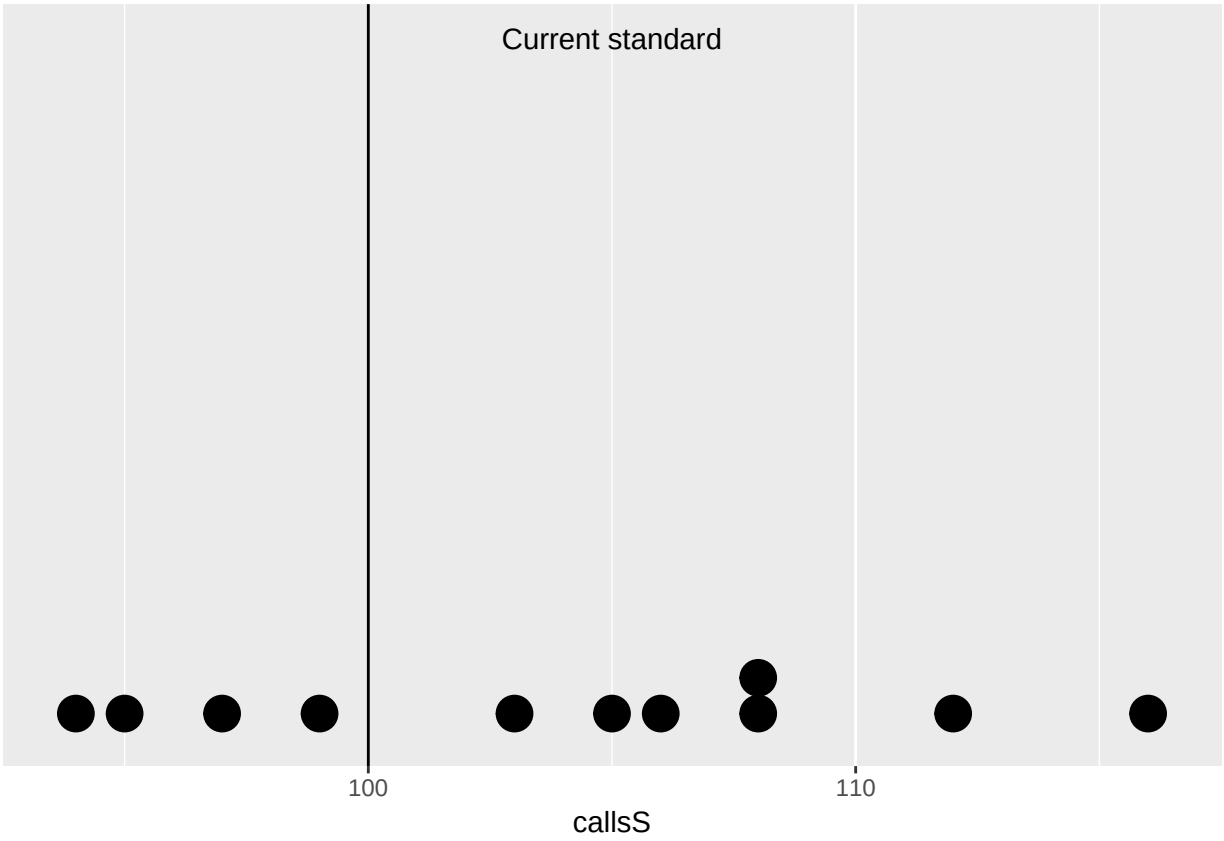
```
# The mean is good. Now lets see

t.test(callsS, mu = 100, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  callsS
## t = 1.8319, df = 10, p-value = 0.04844
## alternative hypothesis: true mean is greater than 100
## 95 percent confidence interval:
```

```
##  100.0414      Inf
## sample estimates:
## mean of x
##  103.9091
```

```
ggplot(df1) +geom_dotplot(aes(x = callsS)) +
  scale_y_continuous(NULL, breaks = NULL) + geom_vline(xintercept = 100) +
  annotate("text", x = 105, y = 4, label = "Current standard")
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# Results Syed did significantly better than the previous standard
```

```
# Example 2

# At another call center Emily made the same claim as Syed,
# and produced the same output, but this time the company does NOT know the distribution
# of the calls, so a nonparametric distribution is used.
callsE = c(112, 99, 105, 103, 108, 95, 116, 97, 106, 108, 94)

wilcox.test(callsE, mu = 100, alternative = "greater")
```

```
## Warning in wilcox.test.default(callsE, mu = 100, alternative = "greater"):
## cannot compute exact p-value with ties
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
```

```
## data:  callsE
## V = 51.5, p-value = 0.0544
## alternative hypothesis: true location is greater than 100
```

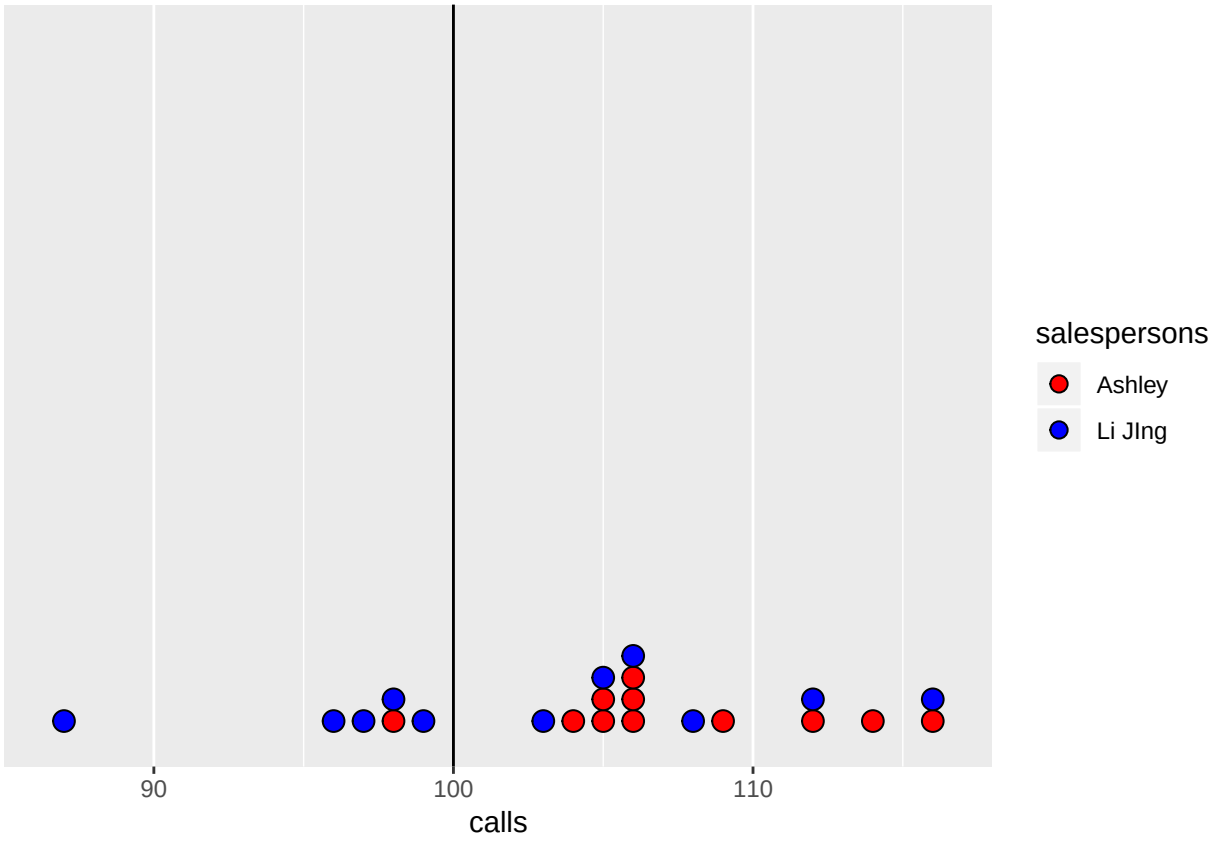
```
# Results: So Syed's results are signficant and Emily's are not
```

```
# Example 3

# Li Jing and Ashley are salespersons. One of them will get promoted based on better performance over
# a few weeks

callsL = c(112, 99, 105, 103, 98, 96, 116, 97, 106, 108, 87)
callsA = c(98, 114, 106, 109, 105, 106, 116, 104, 106, 112, 105)
t.test(callsL, callsA)
```

```
##
##  Welch Two Sample t-test
##
## data:  callsL and callsA
## t = -1.6926, df = 16.742, p-value = 0.109
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.035323   1.217141
## sample estimates:
## mean of x mean of y
##  102.4545  107.3636
```

```
df3 = data.frame(calls = c(callsL, callsA), salespersons = c(rep("Li JIng", 11), rep("Ashley", 11)))

ggplot(df3) +
  geom_dotplot(aes(x = calls, fill = salespersons), dotsize = 0.75, stackgroups = TRUE) +
  scale_y_continuous(NULL, breaks = NULL) + geom_vline(xintercept = 100) +
  scale_fill_manual(values = c("red", "blue"))
```

```
## geom_dotplot called with stackgroups=TRUE and method="dotdensity". You probably want to set binposit
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# Results: Although Ashley scored about 5 points higher than Li Jing, the difference is not significant
```

```
# Example 4

# Now assume for the sake of example, that the data was paired off, so that the paired t-rest would be

t.test(callsL, callsA, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  callsL and callsA
## t = -1.9044, df = 10, p-value = 0.08599
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.6526037   0.8344219
## sample estimates:
## mean of the differences
##               -4.909091
```

```
# Results: Still not significant, but the p value is smaller
```

```
# Example 5

# Marketer Anjali believes that using Calibri font instead of the company standard Arial font
# for an ad will improve the click through rate above the current 2%.
# Some randomly chosen customers were shown Calibri ads with the following results
```

```r
# Total numbers of views = 500
# Number of clicks = 16

binom.test(x = 16, n = 500, p = 0.02, alternative = "greater")
```

```
##
##  Exact binomial test
##
## data:  16 and 500
## number of successes = 16, number of trials = 500, p-value = 0.047
## alternative hypothesis: true probability of success is greater than 0.02
## 95 percent confidence interval:
##  0.02017275 1.00000000
## sample estimates:
## probability of success
##                  0.032
```

```r
# Results: The result is just barely signficant at the 5% level
```

```r
# Example 6

# Marketers Jessica and Michael are trying to get more clicks. Jessica believes a blue font is
# the best choice while Michael believes a red font would work better.
# Potential customers were randomly shown one or the other

# Jessica: 38 clicks out of 1100 views
# Michael 23 clicks out of 1000 views

prop.test(x = c(38, 23), n = c(1100, 1000), alternative = "two.sided")
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(38, 23) out of c(1100, 1000)
## X-squared = 2.0832, df = 1, p-value = 0.1489
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.003649691  0.026740600
## sample estimates:
##     prop 1     prop 2
## 0.03454545 0.02300000
```

```r
# Results: AlthoughJeessica out-performed Michael with 3.6% CTR versus 2.3% CTR,
# the difference was not signficant.
```

```r
# Example 7

# Bian and Liu are trying to market to businesses. Bian believes that the region that the company is in
# and the size of the company are independent attributes. Liu disagrees. They look at the data,
# and find the following
Size = factor(rep(c("Small", "Medium", "Large"),each = 4), levels = c("Small", "Medium", "Large"), orde

df7 = data.frame(Region = rep(c("East", "North", "South", "West"), times = 3),
                 Size = Size,
```

```
                Sales = c(10, 15, 20, 25, 15, 28, 35, 52, 30, 60, 65, 120))
xtabs(Sales ~ Region + Size, df7)
```

```
##        Size
## Region  Small Medium Large
##   East     10     15    30
##   North    15     28    60
##   South    20     35    65
##   West     25     52   120
```

```
M <- as.table(rbind(c(762, 327, 468), c(484, 239, 477)))
dimnames(M) <- list(gender = c("F", "M"),
                    party = c("Democrat","Independent", "Republican"))

df7 = as.table(cbind(c(10, 15, 20, 10), c(15, 28, 35, 72), c(30, 60, 65, 120)))
dimnames(df7) = list(Region = c("East", "North", "South", "West"), Size = c("Small", "Medium", "Large"))
df7
```

```
##        Size
## Region  Small Medium Large
##   East     10     15    30
##   North    15     28    60
##   South    20     35    65
##   West     10     72   120
```

```
chisq.test(df7)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df7
## X-squared = 16.032, df = 6, p-value = 0.01358
```

```
# From the chi squared test for independence, it looks like Liu is correct. The two attributes are
# NOT independent.
```

```
# Example 8

# One-Way ANOVA

# Jennifer and Samantha were unsure which colored font would be most effective
# for some key text on their web pages, so
# they randomly assigned red, green, blue, and orange to pages and collected data
# How well did the various colors do?

counts = 30
colors = factor(rep(c("red", "green", "blue", "orange"), each = counts))

set.seed(2019)

sales = round(c(runif(counts, min = 90, max = 110), runif(counts, min = 85, max = 116),
          runif(counts, min = 97, max = 135), runif(counts, min = 92, max = 105)),0)

df8 = data.frame(colors = colors, sales = sales)
tapply(df8$sales, df8$colors, FUN = mean )
```

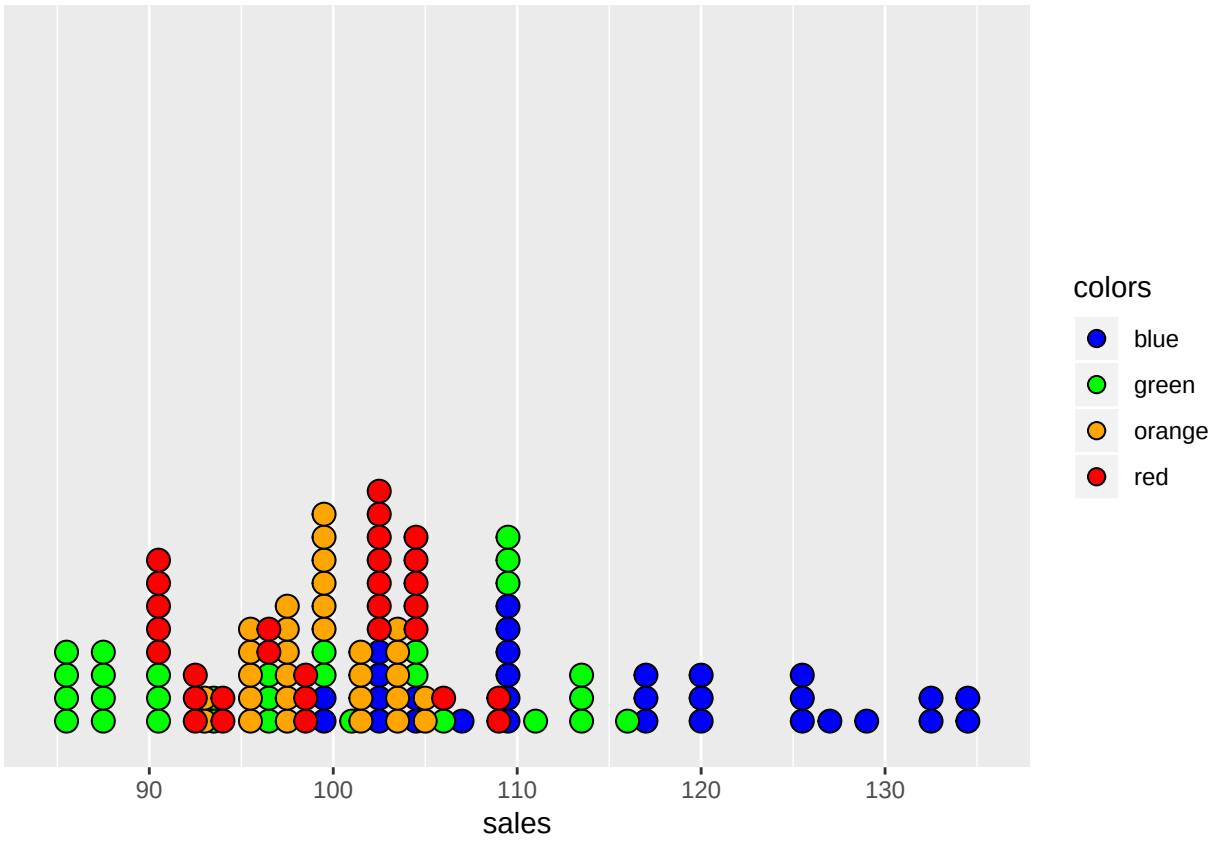
```
##      blue     green    orange       red
## 115.26667  98.40000  99.16667  99.00000
```

```
# It sure looks like there is a big difference between blue and the others.

anova(lm(sales ~ colors, data = df8))
```

```
## Analysis of Variance Table
##
## Response: sales
##            Df Sum Sq Mean Sq F value    Pr(>F)
## colors      3 6069.6 2023.19   28.87 5.123e-14 ***
## Residuals 116 8129.2   70.08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ggplot(df8) +
  geom_dotplot(aes(x = sales, fill = colors), dotsize = 0.75, stackgroups = TRUE) +
  scale_y_continuous(NULL, breaks = NULL) +
  scale_fill_manual(values = c("blue", "green", "orange", "red"))
```

```
## geom_dotplot called with stackgroups=TRUE and method="dotdensity". You probably want to set binposit
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# Results: So the colors are not equal performers
```

```
# Example 9
```

```r
# Two-way ANOVA
# Actually, their manager decided to test for a second feature, font size at the same time.
# The manager used 14, 16, and 18 point fonts, so now the task is to do a two-way ANOVA.

fonts = sample(x = c("F14", "F16", "F18"), size = 4*counts, replace = TRUE)
df8$fonts = fonts
df9 = df8

anova(lm(sales ~ colors + fonts, data = df9))
```

```
## Analysis of Variance Table
##
## Response: sales
##            Df Sum Sq Mean Sq F value    Pr(>F)
## colors      3 6069.6 2023.19 28.5583 7.578e-14 ***
## fonts       2   53.0   26.50  0.3741    0.6888
## Residuals 114 8076.2   70.84
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Result: Colors remain very significant, but fonts are not significant
```

```r
# Example 10
# Back in example 3, Li Jing did about 5 points better than Ashley,
# yet that failed to produce a signficantly
# better performance. It is too late now, but a sample size calculation should have been performed
# prior to the experiment

# Looking for a 5 point spread
power.t.test(delta = 5, sd = 8, sig.level = 0.05, power = 0.9)
```

```
##
##      Two-sample t test power calculation
##
##               n = 54.77642
##           delta = 5
##              sd = 8
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```r
# Wow! The actual sample size was just 11 for each group, but under these assumptions, to find a 5 poin
# difference a sample size of 55 is needed.
```