Stitch Fix A/B Testing

Rayan Roy

Stitch Fix is an online personal styling service that helps take the stress out of shopping for clothes. As a new client you complete a "style quiz" in which you disclose your personal measurements, sizes, preferred fit, preferred styles, and a budget. You are then matched to a personal styler who handpicks clothing, footwear and accessories to match your unique sizes and tastes. Then, at regular intervals in time, your "Fix" (a five-item shipment) is mailed to you. Upon receiving your Fix you may choose to keep 0-5 items and return (for free) any items that you do not wish to purchase. A non-refundable $20 "styling fee" is charged for each Fix, independent of the number of items you keep. However, this $20 is applied as credit toward any items that you do keep.

In the interests of client satisfaction and inventory management, Stitch Fix would like to minimize the *task-completion time* (TCT), the time (in days) between a client submitting their style quiz, and the client's Fix being ready for shipment. This time is influenced in part by inventory, warehousing, and supply chain issues, but is also influenced by the time it takes the Fix to be curated by the stylist. The data scientists on the Stylist & CX Algorithms team are interested in investigating whether average TCT can be reduced by augmenting or replacing the human stylist with a recommendation algorithm. To investigate this they run an experiment with $m = 3$ conditions:

- Condition 1: Fixes are curated solely by a human stylist
- Condition 2: Fixes are curated solely by a machine learning algorithm
- Condition 3: Fixes are curated by a combination of of human and machine input

One thousand clients are randomized into each of these three conditions, and the task-completion time for each client is measured.

(a) What is the metric of interest and what is the corresponding response variable?

The metric of interest is the average task-completion time (TCT) and the corresponding response variabe is the continuous measurement of the task-completion time for each client.

(b) What is the design factor and what are its levels?

The design factor is method used for curating a "Fix" which has 3 levels namely curation of Fix by a human stylist, curation of Fix by a machine learning algorithm and curation of a Fix by a combination of human stylist and machine learning algorithm.

(c) What constitutes an experimental unit in this experiment?

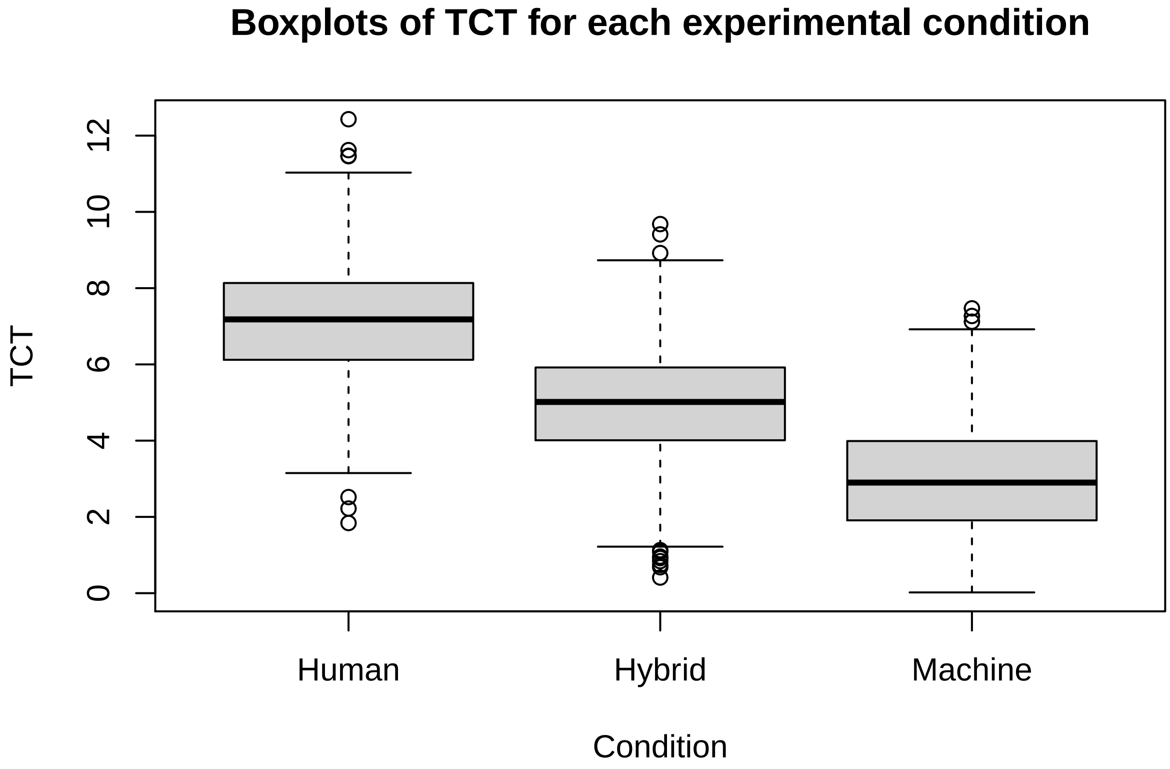
An experimental unit is each of the users of the Stitch Fix's online personal styling service.

(d) State the null and alternative hypotheses for a test of overall equality in the context of this experiment. Use the notation μ_j which represents the metric of interest in condition $j = 1, 2, 3$.

$$H_0 : \mu_1 = \mu_2 = \mu_3 \;\; vs \;\; H_A : \mu_i \neq \mu_j \; for \; i \neq j$$

(e) The file `stitchfix.csv` contains `TCT` measurements for the 3000 clients (1000 in each of the three experimental conditions). Construct side-by-side box plots to visualize the `TCT` distribution in each of the experimental conditions. Comment on which condition appears to minimize task-completion time.

```
sf <- read.csv("stitchfix.csv")
boxplot(sf$TCT ~ sf$Condition, main = "Boxplots of TCT for each experimental condition",
    xlab = "Condition", ylab = "TCT")
```



As seen from the boxplot above, we see that the "Machine" learning algorithm appears to minimize the task-completion time (TCT) since it has the lowest average TCT among the three conditions.

(f) Using the observed data, test the hypothesis in (d) at a 1% significance level. Clearly state your conclusion in the context of the problem, and explain whether this conclusion is surprising, given what you see in the box plots from part (e). **NOTE:** although you may use R to test this hypothesis, be sure to:

- State the formula and the value of the test statistic
- State the null distribution
- State the formula and the value of the p-value

```r
m <- 3
N <- nrow(sf)
cond1 <- sf$TCT[sf$Condition == "Human"]
n1 <- length(cond1)

cond2 <- sf$TCT[sf$Condition == "Machine"]
n2 <- length(cond2)

cond3 <- sf$TCT[sf$Condition == "Hybrid"]
n3 <- length(cond3)

SSC <- n1 * (mean(cond1) - mean(sf$TCT))^2 + n2 * (mean(cond2) - mean(sf$TCT))^2 +
    n3 * (mean(cond3) - mean(sf$TCT))^2
SSC
```

```
## [1] 8663.476
```

```r
SSE <- sum((cond1 - mean(cond1))^2) + sum((cond2 - mean(cond2))^2) + sum((cond3 -
    mean(cond3))^2)
SSE
```

```
## [1] 6489.563
```

```r
t <- (SSC/(m - 1))/(SSE/(N - m))
t
```

```
## [1] 2000.477
```

```r
p_value <- pf(q = t, df1 = m - 1, df2 = N - m, lower.tail = FALSE)
p_value
```

```
## [1] 0
```

The formula and value for test statistic:

$$t = \frac{MSC}{MSE} = \frac{SSC/(m-1)}{SSE/(N-m)} = \frac{8663.476/(3-1)}{6489.563/(3000-3)} = 2000.477$$

The Null distribution is:

$$T \sim F_{(2,2997)}$$

The formula and value for p-value:

$$p = (T \geq t) = P(T \geq 2000.477) = 0$$

Since the p-value < 0.01, we reject the null hypothesis and conclude that at least one of human curated, machine curated or hybrid method has a different average TCT. This is not surprising given that the boxplots show that each of the three conditions have different mean values for TCT.

(g) As a follow-up to part (f), calculate the p-values associated with Student t-tests of each of the following three hypotheses. Note that you need only state the resulting values themselves, and you do not need to justify the use of Student t-tests with F-tests for variances.

$$H_0 : \mu_1 \leq \mu_2 \text{ versus } H_A : \mu_1 > \mu_2$$

```
cond1_1 <- sf$TCT[sf$Condition == "Human"]
cond2_2 <- sf$TCT[sf$Condition == "Machine"]
cond3_3 <- sf$TCT[sf$Condition == "Hybrid"]

test1 <- t.test(x = cond1_1, y = cond2_2, alternative = "greater", mu = 0,
    paired = F, var.equal = T, conf.level = 0.99)
test1
```

```
##
##  Two Sample t-test
##
## data:  cond1_1 and cond2_2
## t = 63.098, df = 1998, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 99 percent confidence interval:
##  4.007804      Inf
## sample estimates:
## mean of x mean of y
##   7.13476   2.97341
```

```
test1$p.value
```

```
## [1] 0
```

The p-value is 0.

$$H_0 : \mu_1 \leq \mu_3 \text{ versus } H_A : \mu_1 > \mu_3$$

```
test2 <- t.test(x = cond1_1, y = cond3_3, alternative = "greater", mu = 0,
    paired = F, var.equal = T, conf.level = 0.99)
test2
```

```
##
##  Two Sample t-test
##
## data:  cond1_1 and cond3_3
## t = 33.119, df = 1998, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 99 percent confidence interval:
##  2.0154    Inf
## sample estimates:
## mean of x mean of y
##   7.13476   4.96697
```

```
test2$p.value
```

```
## [1] 2.062764e-192
```

The p-value is $2.0627641 \times 10^{-192}$

$$H_0 : \mu_2 \geq \mu_3 \text{ versus } H_A : \mu_2 < \mu_3$$

```
test3 <- t.test(x = cond2_2, y = cond3_3, alternative = "less", mu = 0,
    paired = F, var.equal = T, conf.level = 0.99)
test3
```

```
##
##  Two Sample t-test
##
## data:  cond2_2 and cond3_3
## t = -30.197, df = 1998, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 99 percent confidence interval:
##       -Inf -1.839853
## sample estimates:
## mean of x mean of y
##   2.97341   4.96697
```

```
test3$p.value
```

```
## [1] 1.233966e-165
```

The p-value is $1.2339659 \times 10^{-165}$

(h) Using the p-values from (g), in this question you will identify the condition that minimizes average task-completion time while controlling the family-wise error rate. Note that you may use `p.adjust()` where appropriate.

i. [2 points] Calculate the Bonferroni-adjusted p-values and draw your conclusion assuming we wish to ensure $FWER \leq 0.01$.
ii. [2 points] Calculate the Šidák-adjusted p-values and draw your conclusion assuming we wish to ensure $FWER \leq 0.01$.
iii. [2 points] Calculate the Holm-adjusted p-values and draw your conclusion assuming we wish to ensure $FWER \leq 0.01$.

The p-values from part (g) are as follows:

```
p <- c(test1$p.value, test2$p.value, test3$p.value)
M <- length(p)
```

i. Calculate the Bonferroni-adjusted p-values and draw your conclusion assuming we wish to ensure $FWER \leq 0.01$.

Under the Bonferroni method, we get the adjusted p-values:

```
p.adjust(p, method = "bonferroni")
```

```
## [1]  0.000000e+00 6.188292e-192 3.701898e-165
```

Hence we see that p_1, p_2, p_3 are all less than $0.01/3 = 0.0033$. Thus, we reject $H_{0,1},\ H_{0,2}\ and\ H_{0,3}$ while ensuring we have $FWER \leq \alpha^* = 0.01$ for the Bonferroni method.

ii. Calculate the Šidák-adjusted p-values and draw your conclusion assuming we wish to ensure $FWER \leq 0.01$.

Under the Šidák method, we get:

```
1 - (1 - p)^M
```

```
## [1] 0 0 0
```

We see that p_1, p_2, p_3 are all less than $1 - (1 - 0.01)^{1/3} = 0.003344507$. Thus, we reject $H_{0,1},\ H_{0,2}\ and\ H_{0,3}$ while ensuring we have $FWER \leq \alpha^* = 0.01$.

iii. Calculate the Holm-adjusted p-values and draw your conclusion assuming we wish to ensure $FWER \leq 0.01$.

Under the Holm-adjusted method, we get

```
p.adjust(p, method = "holm")
```

```
## [1]  0.000000e+00 4.125528e-192 1.233966e-165
```

We see that $p_{(1)} = p_1 < 0.01/3 = 0.0033$, $p_{(2)} = p_2 < 0.01/2 = 0.005$, and $p_{(3)} = p_3 < 0.01/1 = 0.01$. Thus, we reject $H_{0,1},\ H_{0,2}\ and\ H_{0,3}$ while ensuring we have $FWER \leq \alpha^* = 0.01$.

All the three testing techniques provide the same conclusion. So we can make the conclusion based on hypothesis tests conducted in part (g): The average TCT under human (μ_1) is greater than average TCT under machine (μ_2) and under a hybrid of both human and machine (μ_3). In addition, we also see the average TCT under machine (μ_2) is lower than the hybrid condition (μ_3). Therefore we know that Machine algorithm has the lowest mean TCT while Human Stylist has the highest mean TCT. This is also what we saw in from the boxplot.

(i) Stitch Fix considers a Fix to be "successful" if the majority (≥ 3) of its items are kept. Maximizing *Fix success rate* (FSR), the proportion of Fixes for which 3 or more items are kept, is of interest. Like the task-completion time, Fix success rate is believed to be influenced by whether the Fix is curated by a human stylist, a recommendation algorithm, or a hybrid of the two. To investigate this, in the same experiment described above, the data scientists also recorded a binary indicator for each client which takes the value 1 if the client's Fix was a success, and 0 otherwise. State the null and alternative hypotheses for a test of overall equality of FSRs across the three conditions. Use the notation π_j which represents the FSR in condition $j = 1, 2, 3$.

$$H_0 : \pi_1 = \pi_2 = \pi_3 \quad vs \quad H_A : \pi_i \neq \pi_j \ for \ i \neq j$$

(j) Using the `Fix.Success` data in the `stitchfix.csv` file, test the hypothesis in (i) at a 1% significance level. **NOTE:** although you may use R to test this hypothesis, be sure to:

- State the formula and the value of the test statistic
- State the null distribution
- State the formula and the value of the p-value

```r
cond1_new <- sf$Fix.Success[sf$Condition == "Human"]
n1 <- length(cond1_new)
cond2_new <- sf$Fix.Success[sf$Condition == "Machine"]
n2 <- length(cond2_new)
cond3_new <- sf$Fix.Success[sf$Condition == "Hybrid"]
n3 <- length(cond3_new)

l1 <- length(which(cond1_new == "1"))
l2 <- length(which(cond2_new == "1"))
l3 <- length(which(cond3_new == "1"))

prop.test(x = c(l1, l2, l3), n = c(n1, n2, n3), correct = F)
```

```
##
##  3-sample test for equality of proportions without continuity
##  correction
##
## data:  c(l1, l2, l3) out of c(n1, n2, n3)
## X-squared = 57.491, df = 2, p-value = 3.28e-13
## alternative hypothesis: two.sided
## sample estimates:
## prop 1 prop 2 prop 3
##  0.109  0.043  0.142
```

```r
# pvalue
prop.test(x = c(l1, l2, l3), n = c(n1, n2, n3), correct = F)$p.value
```

```
## [1] 3.280388e-13
```

The formula and value for test-statistic:

$$t = \sum_{i=0}^{1}\sum_{j=1}^{3}\frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} = 57.491$$

The null distribution is:

$$T \sim \chi^2_{(2)}$$

The formula and value for p-value:

$$p = P(T \geq t) = P(T \geq 57.491) = 3.28 * 10^{-13}$$

(k) As a follow-up to part (j), calculate the p-values associated with χ^2-tests of each of the following three hypotheses. Note that you need only state the resulting values themselves.

$$H_0 : \pi_1 \leq \pi_2 \text{ versus } H_A : \pi_1 > \pi_2$$

```
chi_tst1 <- prop.test(x = c(l1, l2), n = c(n1, n2), alternative = "greater",
    correct = F)
chi_tst1
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(l1, l2) out of c(n1, n2)
## X-squared = 31.015, df = 1, p-value = 1.28e-08
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.04665845 1.00000000
## sample estimates:
## prop 1 prop 2
##  0.109  0.043
```

```
chi_tst1$p.value
```

```
## [1] 1.280185e-08
```

The p-value is 1.2801851×10^{-8}.

$$H_0 : \pi_1 \geq \pi_3 \text{ versus } H_A : \pi_1 < \pi_3$$

```
chi_tst2 <- prop.test(x = c(l1, l3), n = c(n1, n3), alternative = "less",
    correct = F)
chi_tst2
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(l1, l3) out of c(n1, n3)
## X-squared = 4.9613, df = 1, p-value = 0.01296
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.00000000 -0.00866089
## sample estimates:
## prop 1 prop 2
##  0.109  0.142
```

```
chi_tst2$p.value
```

```
## [1] 0.01296045
```

The p-value is 0.0129605.

$$H_0 : \pi_2 \geq \pi_3 \text{ versus } H_A : \pi_2 < \pi_3$$

```
chi_tst3 <- prop.test(x = c(l2, l3), n = c(n2, n3), alternative = "less",
    correct = F)
chi_tst3
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(l2, l3) out of c(n2, n3)
## X-squared = 58.378, df = 1, p-value = 1.081e-14
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.00000000 -0.07800075
## sample estimates:
## prop 1 prop 2
##  0.043  0.142
```

```
chi_tst3$p.value
```

```
## [1] 1.081273e-14
```

The p-value is $1.0812733 \times 10^{-14}$.

(l) Using the p-values from (k), in this question you will identify the condition that maximizes Fix success rate while accounting for the multiple comparison problem. Note that you may use `p.adjust()` where appropriate.

  i. [2 points] Calculate the Bonferroni-adjusted p-values and draw your conclusion assuming we wish to ensure $FWER \leq 0.01$.
  ii. [2 points] Calculate the Holm-adjusted p-values and draw your conclusion assuming we wish to ensure $FWER \leq 0.01$.
  iii. [2 points] Calculate the Benjamini-Hochberg-adjusted p-values and draw your conclusion assuming we wish to ensure $FDR \leq 0.01$.

From part (k), the p-values we get are:

```
p_val <- c(chi_tst1$p.value, chi_tst2$p.value, chi_tst3$p.value)
M_new <- length(p_val)
```

i. Calculate the Bonferroni-adjusted p-values and draw your conclusion assuming we wish to ensure $FWER \leq 0.01$.

```
p.adjust(p_val, method = "bonferroni")
```

```
## [1] 3.840555e-08 3.888136e-02 3.243820e-14
```

We see that p_1, p_3 are all less than $0.01/3 = 0.0033$, but p_2 is greater than 0.0033. Thus, we reject $H_{0,1}$ *and* $H_{0,3}$ and fail to reject $H_{0,2}$.

ii. Calculate the Holm-adjusted p-values and draw your conclusion assuming we wish to ensure $FWER \leq 0.01$.

```
p.adjust(p_val, method = "holm")
```

```
## [1] 2.560370e-08 1.296045e-02 3.243820e-14
```

We see that $p_{(1)} = p_3 = 3.243820e - 14 < 0.01/3 = 0.0033$, $p_{(2)} = p_1 = 2.560370e - 08 < 0.01/2 = 0.005$, and $p_{(3)} = p_2 = 0.01296 > 0.01/1 = 0.010$. Thus, we reject $H_{0,(1)} = H_{0,3}$ *and* $H_{0,(2)} = H_{0,1}$ and fail to reject $H_{0,(3)} = H_{0,2}$.

iii. Calculate the Benjamini-Hochberg-adjusted p-values and draw your conclusion assuming we wish to ensure $FDR \leq 0.01$.

```
p.adjust(p_val, method = "BH")
```

```
## [1] 1.920278e-08 1.296045e-02 3.243820e-14
```

We see that $p_{(1)} = p_3 = 3.243820e - 14 < 1 * 0.01/3 = 0.0033$, $p_{(2)} = p_1 = 1.920278e - 08 < 2 * 0.01/3 = 0.0066$, and $p_{(3)} = p_2 = 0.01296 > 3 * 0.01/3 = 0.01$. Thus, we reject $H_{0,1}$ *and* $H_{0,3}$ and fail to reject $H_{0,2}$.

All the three testing techniques provide the same conclusion (i.e $\pi_1 \geq \pi_3 > \pi_2$). We can make the conclusion based on hypothesis tests conducted in part (k): The FSR under human stylist (π_1) is greater than FSR under machine (π_2) and the FSR under machine (π_2) is lower than FSR under hybrid approach (π_3). However, we also see the FSR under human is greater than or equal to FSR under hybrid condition indicating that either of human or hybrid condition maximizes FSR. Therefore, ensuring $FWER \leq 0.01$ in first two parts of (l) and $FDR \leq 0.01$ in third part of (l), we know the fix success rate is maximized by Fixes that are either created by a human stylist or an hybrid approach of human and ML algorithm.