Case Study Appendix

Colin Pi, Sharan Ganjam Seshachallam

2019 2 25

Data Wrangling

```
# Import the Dataset
kick.2018 <- read_csv("ks-projects-201801.csv")

# Sort out failed and successful projects
kick.2018 <- kick.2018 %>% filter(state %in% c("failed", "successful")) %>%
    mutate(diff_date = as.numeric(as.Date(deadline) - as.Date(str_extract(launched,
        "^.{10}"))))

kick.2018$state <- factor(kick.2018$state)

# Random Sampling of the Data

## index <- sample(nrow(kick.2018), 2000, replace=FALSE)

## kick.sample <- kick.2018[index, ]

## write.csv(kick.sample, file = 'Kickstarter_sample.csv', row.names = FALSE)
```

Launched-Deadline and Success/Failure

Summary Statistic

```
kickstarter <- read_csv("Kickstarter_sample.csv")  ## pull in sample data

failed <- subset(kickstarter, select = diff_date, subset = state == "failed",
    drop = T)  ## subset failed projects
successful <- subset(kickstarter, select = diff_date, subset = state == "successful",
    drop = T)  ## subset successful projects

stargazer(data.frame(kickstarter$diff_date), header = FALSE, title = "Summary statistics of the duratio
```

Table 1: Summary statistics of the duration of investment window (pooled data, days)

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
kickstarter.diff_date	2,000	34.068	13.093	2	30	36	91

```
stargazer(data.frame(failed), header = FALSE, title = "Summary statistics of the duration of investment
```

```
stargazer(data.frame(successful), header = FALSE, title = "Summary statistics of the duration of invest
```

Table 2: Summary statistics of the duration of investment window (failed, days)

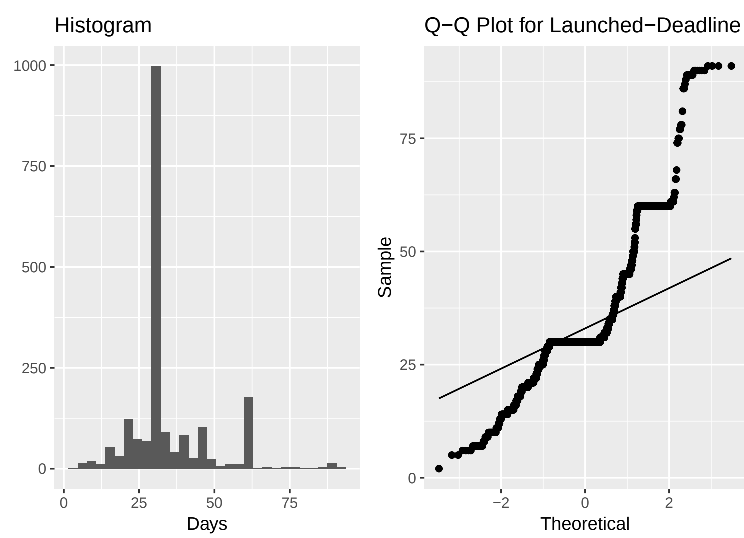
Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
failed	1,174	35.395	13.386	5	30	40	91

Table 3: Summary statistics of the duration of investment window (successful, days)

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
successful	826	32.182	12.432	2	29	34.8	91

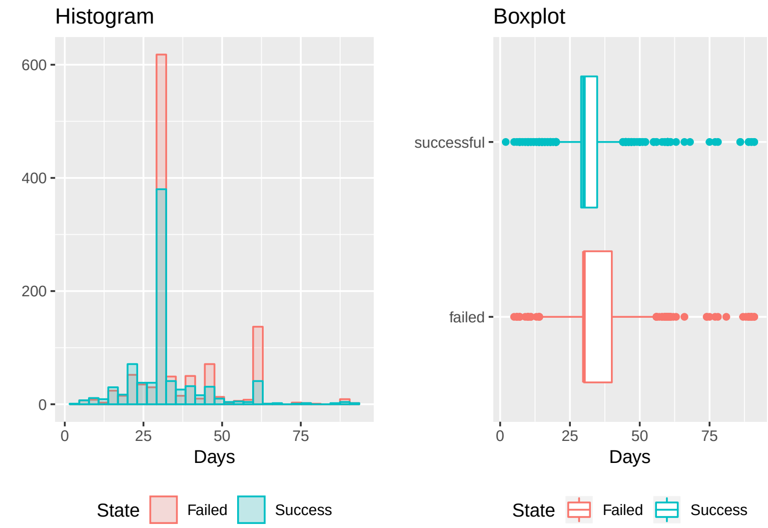
Data Visualization

```
pooled.hist <- ggplot(kickstarter) + geom_histogram(aes(x = diff_date), bins = 30) +
    xlab("Days") + ylab("") + ggtitle("Histogram")

kickstarter$state <- factor(kickstarter$state)

pooled.qqplot <- ggplot(kickstarter, aes(sample = diff_date)) + stat_qq() +
    stat_qq_line() + xlab("Theoretical") + ylab("Sample") + ggtitle("Q-Q Plot for Launched-Deadline")

grid.arrange(pooled.hist, pooled.qqplot, ncol = 2)
```

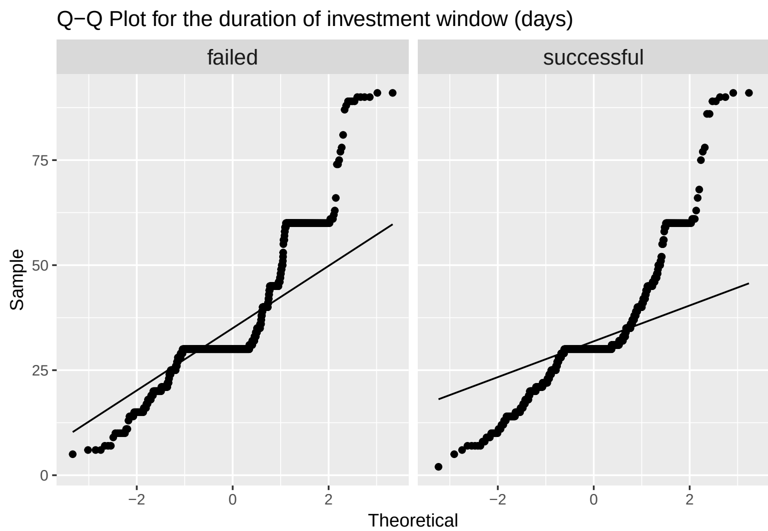



```
category.hist <- ggplot(kickstarter, aes(x = diff_date, fill = state, color = state)) +
    geom_histogram(alpha = 0.2, position = "identity", bins = 30) + xlab("Days") +
    ylab("") + scale_fill_discrete(name = "State", breaks = c("failed", "successful"),
    labels = c("Failed", "Success")) + scale_color_discrete(name = "State",
    breaks = c("failed", "successful"), labels = c("Failed", "Success")) + ggtitle("Histogram") +
    theme(legend.position = "bottom")

category.box <- ggplot(kickstarter, aes(x = state, y = diff_date, color = state)) +
    geom_boxplot() + xlab("") + ylab("Days") + scale_color_discrete(name = "State",
    breaks = c("failed", "successful"), labels = c("Failed", "Success")) + ggtitle("Boxplot") +
    theme(legend.position = "bottom") + coord_flip()
```

```r
grid.arrange(category.hist, category.box, ncol = 2)
```




```r
ggplot(kickstarter, aes(sample = diff_date)) + stat_qq() + stat_qq_line() +
    xlab("Theoretical") + ylab("Sample") + ggtitle("Q-Q Plot for the duration of investment window (day
    facet_wrap(~state) + theme(strip.text.x = element_text(size = 13))
```



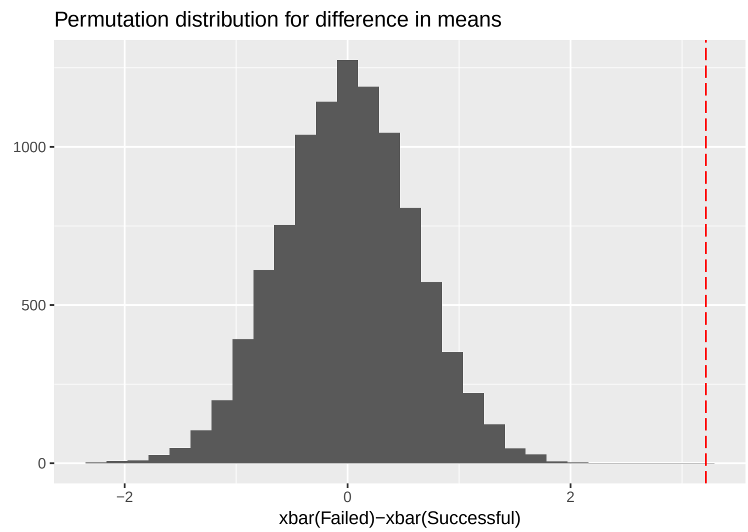
Hypothesis Testing

$$H_0 : \mu_{investment\ window,\ failed} = \mu_{investment\ window,\ successful}$$

$$H_\alpha : \mu_{investment\ window,\ failed} \neq \mu_{investment\ window,\ successful}$$

```r
mean.diff <- mean(failed)-mean(successful)
N <- 10^4
pooled.data <- unlist(kickstarter$diff_date)
perm <- numeric(N)

for (i in 1:N){
  index <- sample(length(pooled.data), size = length(failed), replace = FALSE)
```

```r
  perm[i] <- mean(pooled.data[index])-mean(pooled.data[-index])
}

ggplot(data.frame(perm), aes(x = perm)) +
  geom_histogram(bins = 30) +
  xlab("xbar(Failed)-xbar(Successful)") +
  ylab("") +
  geom_vline(xintercept=mean.diff, color = "red", linetype = "longdash") +
  ggtitle("Permutation distribution for difference in means")
```




```r
p.perm <- (sum(perm >= mean.diff)+1)/(N+1)*2
t.test(diff_date~state, data = kickstarter)
```

```
##
##  Welch Two Sample t-test
##
## data:  diff_date by state
## t = 5.5135, df = 1852.8, p-value = 4.012e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.070483 4.356781
## sample estimates:
##     mean in group failed mean in group successful
##                 35.39523                 32.18160
```

p-value (perm): $1.9998 \cdot 10^{-4}$
p-value (t): $4.0121047 \cdot 10^{-8}$

Confidence Interval

```r
se <- sqrt(var(failed)/length(failed) + var(successful)/length(successful))

boot.perc <- numeric(N)
Tstar <- numeric(N)

for (i in 1:N) {
```

```r
    failedBoot <- sample(failed, length(failed), replace = TRUE)
    successfulBoot <- sample(successful, length(successful), replace = TRUE)
    SEstar <- sqrt(var(failedBoot)/length(failedBoot) + var(successfulBoot)/length(successfulBoot))
    meanDiffBoot <- mean(failedBoot) - mean(successfulBoot)
    Tstar[i] <- (meanDiffBoot - mean.diff)/SEstar
    boot.perc[i] <- meanDiffBoot
}

formula.t.CI <- t.test(diff_date ~ state, data = kickstarter)$conf
boot.perc.CI <- quantile(boot.perc, c(0.025, 0.975))
boot.t.CI <- mean.diff - quantile(Tstar, c(0.975, 0.025)) * se
```

Formula t CI (95%): (2.07, 4.357)
Bootstrap percentile CI (95%): (2.093, 4.327)
Bootstrap t CI (95%): (2.093, 4.324)

Hypothsis testing on the population

```r
t.test(diff_date~state, data = kick.2018)
```

```
##
##  Welch Two Sample t-test
##
## data:  diff_date by state
## t = 68.989, df = 307680, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.931166 3.102584
## sample estimates:
##     mean in group failed mean in group successful
##                 35.17332                 32.15645
```

Two-sided hypothesis t-test p-value: $< 2.2 \cdot 10^{-16}$
Formula t CI (95%): (2.931166, 3.102584)