LAB 6

RAJ SHAH

2025-04-30

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

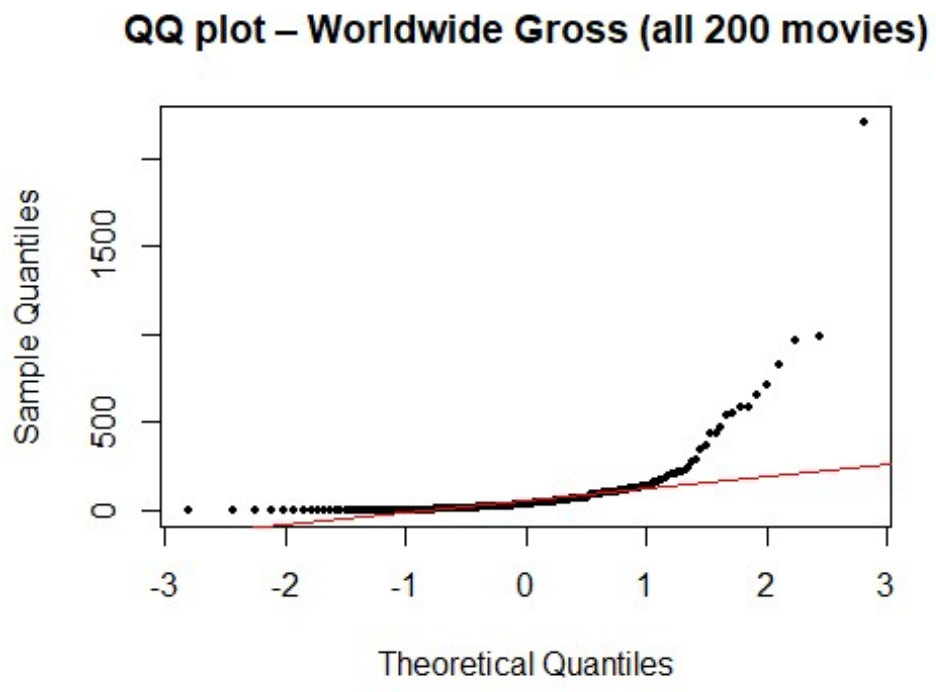
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## ── 1. LOAD & CLEAN ──────────────────────────────────────────
full <- read.csv("C:\\Users\\RAJ RUTGERS\\Downloads\\movie_boxoffice-1.csv",
                 stringsAsFactors = FALSE)

set.seed(1234)
cat("Random seed number:", 1234, "\n")
```

```
## Random seed number: 1234
```

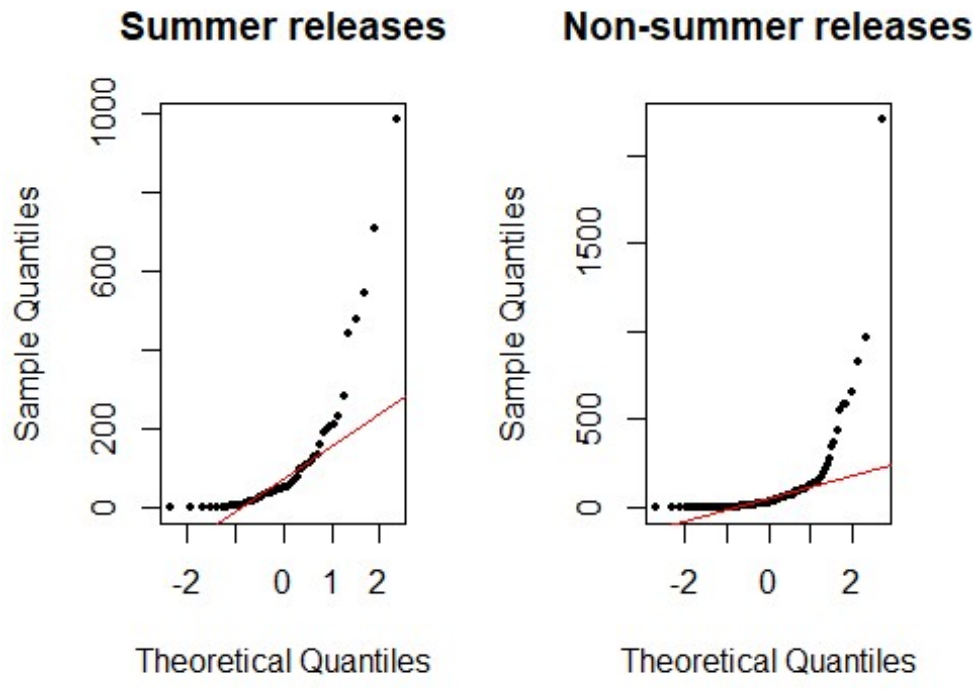
```
movies <- dplyr::sample_n(full, 200)

num_cols <- c("Budget", "Domestic_Gross", "Worldwide_Gross")
movies[num_cols] <- lapply(movies[num_cols],
                           function(x) as.numeric(gsub("[\\$,]", "", x)))

movies <- movies %>%
  mutate(
    over_budget = Worldwide_Gross > Budget,
    is_summer   = Month %in% c("Jun", "Jul", "Aug"),
    era         = case_when(
      Year >= 1980 & Year <= 1999 ~ "1980-1999",
      Year >= 2000 & Year <= 2018 ~ "2000-2018",
      TRUE                        ~ "other")
  )

## ── GLOBAL SETTINGS ──────────────────────────────────────────
set.seed(1234)       # resampling reproducibility
B <- 1e4
n <- nrow(movies)   # 200 rows

interp <- function(p, alt) {
```

```r
  if (p < 0.05) paste("Reject H0 – evidence supports", alt)
  else          paste("Fail to reject H0 – no significant evidence for", alt)
}

## ── Q1: ONE-SAMPLE MEAN (μ = 90) ──────────────────────────────
mu0  <- 90
xbar <- mean(movies$Worldwide_Gross);  s <- sd(movies$Worldwide_Gross)

centered <- movies$Worldwide_Gross - xbar + mu0          # null-centred
boot_means <- replicate(B, mean(sample(centered, n, TRUE)))
p_sim_mean <- mean(abs(boot_means - mu0) >= abs(xbar - mu0))

z_mean      <- (xbar - mu0) / (s / sqrt(n))
p_theor_mean <- 2 * (1 - pnorm(abs(z_mean)))

cat("Q1 – μ = 90\n",
    "      n               :", n, "\n",
    "      Sample mean     :", round(xbar, 2), "\n",
    "      Sample SD       :", round(s, 2),  "\n",
    "   -- Simulation (null-bootstrap) --\n",
    "      p-value         :", round(p_sim_mean, 4), "\n",
    "   -- Theoretical Z test --\n",
    "      Z statistic     :", round(z_mean, 3),  "\n",
    "      p-value         :", round(p_theor_mean,4), "\n",
    "   →", interp(p_theor_mean, "a difference from $90 M"), "\n\n")

## Q1 – μ = 90
##       n               : 200
##       Sample mean     : 100.85
##       Sample SD       : 220.43
##    -- Simulation (null-bootstrap) --
##       p-value         : 0.4882
##    -- Theoretical Z test --
##       Z statistic     : 0.696
##       p-value         : 0.4864
##    → Fail to reject H0 – no significant evidence for a difference from
$90 M

cat("Check: n =", n, "> 30  ⇒ normal approx OK\n\n")

## Check: n = 200 > 30  ⇒ normal approx OK

#The data show no significant difference between the overall mean worldwide
gross and $90 million (p ≈ 0.49).

qqnorm(movies$Worldwide_Gross,
       main = "QQ plot – Worldwide Gross (all 200 movies)",
       pch = 19, cex = 0.6)
qqline(movies$Worldwide_Gross, col = "red")
```




```r
## ── Q2: SUMMER (Jun-Aug) > REST ──────────────────────────────
summer   <- movies %>% filter(is_summer);    n1 <- nrow(summer)
nonsummer<- movies %>% filter(!is_summer);   n2 <- nrow(nonsummer)

xbar1 <- mean(summer$Worldwide_Gross);  s1 <- sd(summer$Worldwide_Gross)
xbar2 <- mean(nonsummer$Worldwide_Gross); s2 <- sd(nonsummer$Worldwide_Gross)

diff_obs <- xbar1 - xbar2
perm_diffs <- replicate(B, {
  sh <- sample(movies$Worldwide_Gross)
  mean(sh[1:n1]) - mean(sh[(n1+1):n])
})
p_sim_diff <- mean(perm_diffs >= diff_obs)

se_diff <- sqrt((s1^2)/n1 + (s2^2)/n2)
z_diff  <- diff_obs / se_diff
p_theor_diff <- 1 - pnorm(z_diff)

cat("Q2 – Summer > Rest\n",
    "      n_summer        :", n1, "\n",
    "      n_rest          :", n2, "\n",
    "      Mean summer     :", round(xbar1, 2), "\n",
    "      Mean rest       :", round(xbar2, 2), "\n",
    "      Observed diff   :", round(diff_obs, 2), "\n",
    "   -- Simulation (permutation) --\n",
    "      p-value         :", round(p_sim_diff,4), "\n",
```

```
    "   -- Theoretical Z test --\n",
    "      Z statistic     :", round(z_diff, 3), "\n",
    "      p-value         :", round(p_theor_diff,4), "\n",
    "   →", interp(p_theor_diff, "higher summer earnings"), "\n\n")

## Q2 – Summer > Rest
##        n_summer        : 53
##        n_rest          : 147
##        Mean summer     : 121.04
##        Mean rest       : 93.57
##        Observed diff   : 27.47
##     -- Simulation (permutation) --
##        p-value         : 0.2154
##     -- Theoretical Z test --
##        Z statistic     : 0.85
##        p-value         : 0.1977
##     → Fail to reject H0 – no significant evidence for higher summer
earnings

#Summer releases do not earn significantly more than non-summer releases in
this sample (p ≈ 0.20).

par(mfrow = c(1, 2))  # two plots side-by-side

qqnorm(summer$Worldwide_Gross,
       main = "Summer releases",
       pch = 19, cex = 0.6)
qqline(summer$Worldwide_Gross, col = "red")

qqnorm(nonsummer$Worldwide_Gross,
       main = "Non-summer releases",
       pch = 19, cex = 0.6)
qqline(nonsummer$Worldwide_Gross, col = "red")
```




```r
par(mfrow = c(1, 1))

## ── Q3: ONE-SAMPLE PROPORTION (p = 0.70) ──────────────────────
p0   <- 0.70
phat <- mean(movies$over_budget)

boot_props <- replicate(B, mean(rbinom(n, 1, p0)))     # null-resample
p_sim_prop <- mean(abs(boot_props - p0) >= abs(phat - p0))

z_prop       <- (phat - p0) / sqrt(p0*(1-p0)/n)
p_theor_prop <- 2 * (1 - pnorm(abs(z_prop)))

cat("Q3 – p ≠ 0.70\n",
    "      n                :", n, "\n",
    "      phat             :", round(phat, 3), "\n",
    "   -- Simulation (null resample) --\n",
    "      p-value          :", round(p_sim_prop,4), "\n",
    "   -- Theoretical Z test --\n",
    "      Z statistic      :", round(z_prop ,3), "\n",
    "      p-value          :", round(p_theor_prop,4), "\n",
    "   →", interp(p_theor_prop, "a proportion different from 0.70"), "\n\n")

## Q3 – p ≠ 0.70
##        n                : 200
##        phat             : 0.62
##     -- Simulation (null resample) --
```

```
##        p-value         : 0.0154
##     -- Theoretical Z test --
##        Z statistic     : -2.469
##        p-value         : 0.0136
##     → Reject H0 – evidence supports a proportion different from 0.70
```

```
tab3 <- table(movies$over_budget)
cat("Check: successes =", tab3["TRUE"],
    ", failures =", tab3["FALSE"], "  (both ≥ 10)\n\n")
```

```
## Check: successes = 124 , failures = 76   (both ≥ 10)
```

```
#About 62 % of movies beat their budget—significantly lower than the
hypothesized 70 % (p ≈ 0.015).
```

```
## ── Q4: ERA 80-99 < 00-18 (proportions) ──────────────────────
era1 <- movies %>% filter(era == "1980-1999");  n1p <- nrow(era1); p1 <-
mean(era1$over_budget)
era2 <- movies %>% filter(era == "2000-2018");  n2p <- nrow(era2); p2 <-
mean(era2$over_budget)
diff_obs_p <- p1 - p2

perm_diff_p <- replicate(B, {
  sh <- sample(movies$over_budget)
  mean(sh[1:n1p]) - mean(sh[(n1p+1):(n1p+n2p)])
})
p_sim_era <- mean(perm_diff_p <= diff_obs_p)

p_pool <- (sum(era1$over_budget)+sum(era2$over_budget))/(n1p+n2p)
se_pool <- sqrt(p_pool*(1-p_pool)*(1/n1p + 1/n2p))
z_era   <- diff_obs_p / se_pool
p_theor_era <- pnorm(z_era)

cat("Q4 – Era 80-99 < 00-18\n",
    "      n_80-99         :", n1p, "\n",
    "      n_00-18         :", n2p, "\n",
    "      p̂ 80-99         :", round(p1, 3), "\n",
    "      p̂ 00-18         :", round(p2, 3), "\n",
    "      Observed diff   :", round(diff_obs_p, 3), "\n",
    "   -- Simulation (permutation) --\n",
    "      p-value         :", round(p_sim_era,4), "\n",
    "   -- Theoretical Z test --\n",
    "      Z statistic     :", round(z_era ,3), "\n",
    "      p-value         :", round(p_theor_era,4), "\n",
    "   →", interp(p_theor_era, "a lower 1980-99 proportion"), "\n")
```

```
## Q4 – Era 80-99 < 00-18
##        n_80-99         : 53
##        n_00-18         : 147
##        p̂ 80-99         : 0.698
```

```
##        p̂ 00-18         : 0.592
##        Observed diff   : 0.106
##     -- Simulation (permutation) --
##        p-value         : 0.9402
##     -- Theoretical Z test --
##        Z statistic     : 1.367
##        p-value         : 0.9141
##     → Fail to reject H0 – no significant evidence for a lower 1980-99
proportion

#The 1980-99 beat-budget rate is not lower than the 2000-18 rate (p ≈ 0.91);
we fail to reject H₀.
```