Activity to explore multiple comparisons (sample solution)

NAME HERE"

2025-05-24

The perils of multiple comparison

During the course, we have learned about some of the perils of *dubious practices* such as overfitting models, making errors in extrapolation, and ignoring or misinterpreting null results. While there is not a single way to conduct an analysis, researchers need to be transparent in their decision-making process and acknowledge the limitations behind their research.

Today you will be working in pairs to explore the potential problems that arise when care isn't taken to account for multiple comparisons.

Question 1

A statistical analyst carried out an investigation of the association of gender and teaching evaluations at a university. They undertook exploratory analysis of the data and carried out a number of bivariate comparisons. The multiple items on the teaching evaluation were consolidated to a single measure based on these exploratory analyses. They used this information to construct a multivariable regression model that found evidence for biases.

What issues might arise based on such an analytic approach?

SOLUTION:

The use of the observed data to select the predictors included in the multivariable model will tend to inflate Type-I error rates. Use of a holdout sample (or other cross-validation approach) would be necessary to avoid anti-conservative inferences.

Question 2

The team wants assess how bad the potential is for inflation of Type-I errors using their approach Imagine that they have 100 predictor variables and one outcome measured for $n = 250$ observations).

The investigators use the following procedure:

1. Fit each of the 100 bivariate models for the outcome as a function of a single predictor, then
2. Include all of the significant predictors in the overall model.

What is the distribution of the p-value for the overall F-test for the MLR model, assuming that there are no associations between any of the predictors and the outcome (all are assumed to be multivariate normal and independent). Carry out a simulation to check your answer.

SOLUTION:

```
numsim <- 1000
set.seed(1998)
genmodel <- function(p = 100, n = 250, alpha = 0.05, twostage = TRUE) {
  X <- matrix(rnorm(p * n), nrow = n)
  y <- rnorm(n)
  keep <- logical(p)
  for (i in 1:p) {
    mod <- lm(y ~ X[, i])
    testpval <- coef(summary(mod))[2, 4]
    keep[i] <- testpval < alpha
  }
  if (twostage == TRUE) {
    keep <- logical(p)
    for (i in 1:p) {
      mod <- lm(y ~ X[, i])
      testpval <- coef(summary(mod))[2, 4]
      keep[i] <- testpval < alpha
    }
    smallX <- data.frame(y, X[, keep])
  } else { # include all predictors
    smallX <- data.frame(y, X)
  }
  overall <- lm(y ~ ., data = smallX)
  return(overallp = broom::glance(overall)$p.value)
}
```

```
tictoc::tic()
res <- tibble(   # should be uniform
  sim = 1:numsim,
  p_value = sim |> map_dbl(~ genmodel(twostage = FALSE))
)
tictoc::toc()
```

```
27.21 sec elapsed
```

```
mosaic::binom.test(~ (p_value < 0.05), data = res)
```

```
data:  res$(p_value < 0.05)  [with success = TRUE]
number of successes = 55, number of trials = 1000, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.04169880 0.07099152
sample estimates:
probability of success
                 0.055
```

When all predictors are included in the model, the resulting overall p-value is uniform over the interval from zero to one, with only about 5% between 0 and 0.05.

```
ggplot(data = res, aes(x = p_value)) +
  geom_histogram(binwidth = 0.05) +
  xlab("distribution of overall p-value")
```

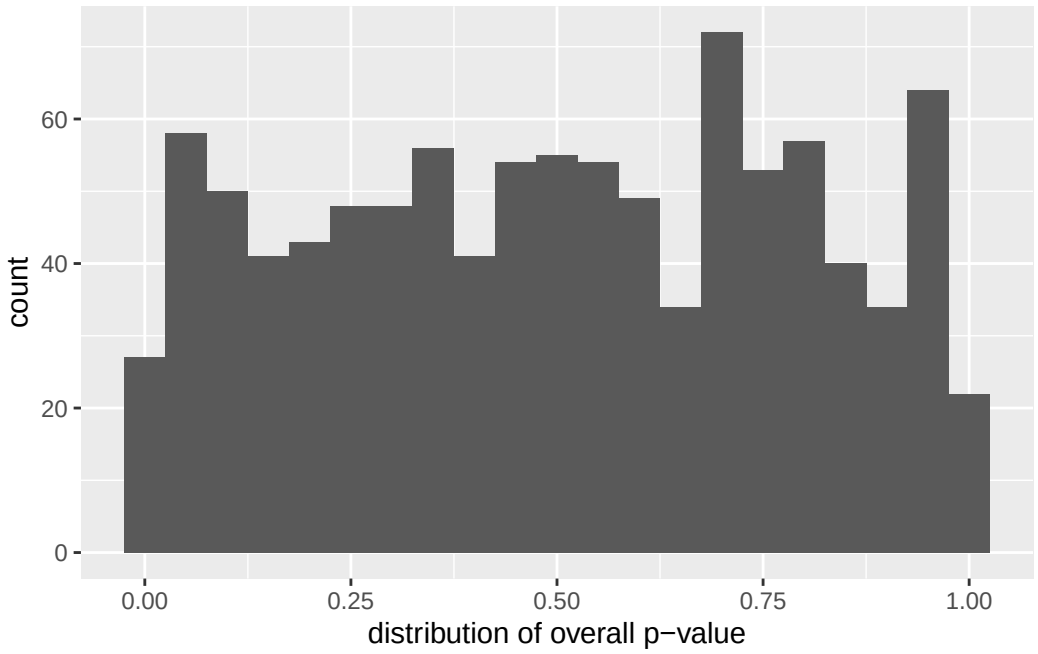

Figure 1: Type-I error rate is maintained

```
tictoc::tic()
res <- res |>
  mutate(
    # no longer uniform
    p_value2 = sim |> map_dbl(~ genmodel(twostage = TRUE))
  )
tictoc::toc()
```

```
44.345 sec elapsed
```

```
mosaic::binom.test(~ (p_value2 < 0.05), data = res)
```

```
data:  res$(p_value2 < 0.05)  [with success = TRUE]
number of successes = 993, number of trials = 993, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.996292 1.000000
```

```
sample estimates:
probability of success
                     1
```

Note that sometimes that model doesn't always converge (if none of the predictors are statistically significant).

```
ggplot(data = res, aes(x = p_value2)) +
  geom_histogram(binwidth = 0.005) +
  xlab("distribution of overall p-value")
```

```
Warning: Removed 7 rows containing non-finite outside the scale range
(`stat_bin()`).
```

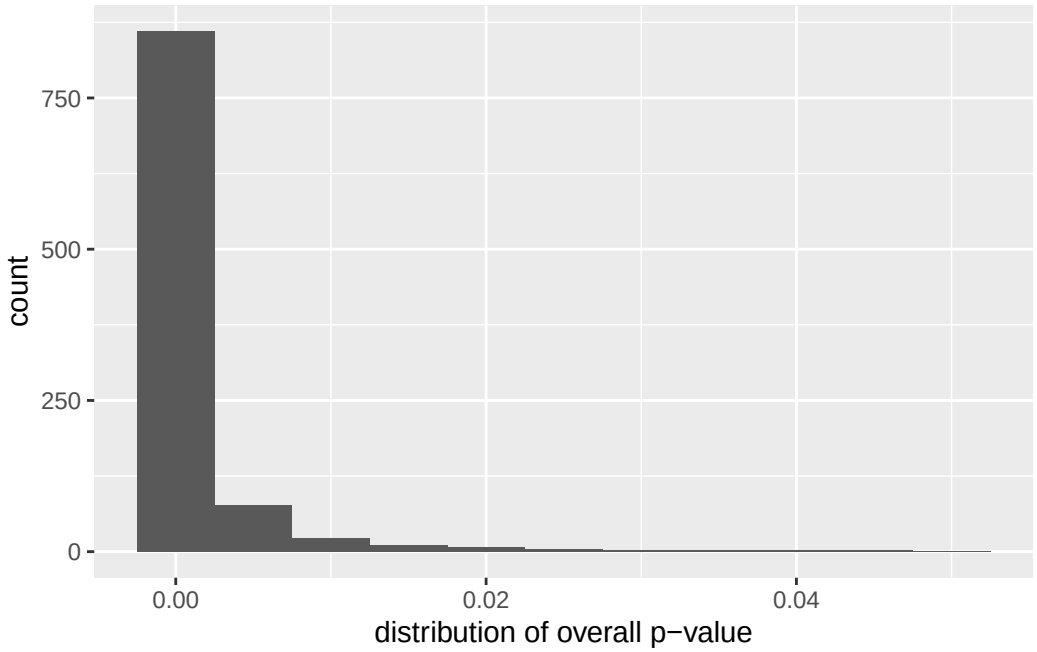

Figure 2: Type-I error rate is not maintained

When only the significant predictors are included, the resulting overall p-values are much more likely to be significant (this procedure leads to a dramatically inflated Type-I error rate).