Distribution and Significance

Max Turgeon

DATA 2010–Tools and Techniques in Data Science

Lecture Objectives

- Explain hypothesis testing, p-values, and multiple testing.
- Perform simple statistical procedures and correctly interpret the results.
- Use R to conduct a permutation test.

Motivation

- In the previous lecture, we discussed correlations.
    - A statistician's favourite way of describing relationships.
- But what is strong correlation? Does it depend on sample size?
- In this lecture, we will review some important concepts related to statistical significance.

Definitions i

First, recall the following definitions:

- A **statistic** is a function of a sample, i.e. from a sample $X_1, \ldots, X_n$ compute an output.
  - Sample mean, sample variance, etc.
  - Histogram, empirical CDF
- An **estimator** is a statistic $\hat{\theta}$ used to estimate (or "approximate") a population parameter θ.
  - The sample mean estimates the population mean.
  - The empirical CDF approximates the population CDF.

Definitions ii

- A statistic is a random variable, because it is a function of the sample. Therefore it has a distribution: the **sampling distribution**.
    - If $X_1, \ldots, X_n$ are $N(\mu, \sigma^2)$, then the sampling distribution for the sample mean is $N(\mu, \sigma^2/n)$

Remark

- The sampling distribution is often a function of unknown population parameters.
    - Or even the type of distribution may be unknown.

Hypothesis testing i

- In hypothesis testing, we start with a **null hypothesis** about our parameter θ:

$$H_0 : \theta = \theta_0.$$

- We then use a **test statistic** to determine whether we should reject or not the null hypothesis.
    - A test statistic can also be an estimator, but more often it's a transformation thereof.
- If we know the sampling distribution of our test statistic when H_0 holds (i.e. $\theta = \theta_0$), then we can compute *how likely* it is to observe some given values of a test statistic.

Hypothesis testing ii

- This gives rise to the notion of a **p-value**: if your test statistic is T, and the observed value (i.e. after you've plugged in your data) is t, then the p-value is the following conditional probability:

$$P(T > t \mid H_0 \text{ hold}).$$

- Finally, we can reject the null hypothesis if the p-value is smaller than a predetermined level of significance α.

Misconceptions about p-values

- P-values are often misinterpreted. The following statements are all **false**:
  - A p-value is the probability the null hypothesis is true.
  - A non-significant test means the null hypothesis is true.
  - Only significant p-values are worthy of interest.
- For this reason, there has been a lot of recent discussions about the utility of p-values, whether they should be replaced by something else, etc.
  - For more information, see *The ASA Statement on p-Values: Context, Process, and Purpose* (American Statistician, 2016)

Motivating example—T test i

- We will use the `chickwts` dataset (available in base `R`).
- Contains 71 observations: chick weight, and the type of feed use.
- We will focus on two types of feed: *soybean* and *linseed*
- First, let's create 95% confidence interval.

Motivating example—T test ii

```
library(tidyverse)
dataset <- chickwts |>
  filter(feed %in% c("soybean", "linseed"))

dataset |>
  group_by(feed) |>
  summarise(samp_mean = mean(weight),
            std_err = sd(weight)/sqrt(n())) |>
  mutate(low_bd = samp_mean - 1.96*std_err,
         upp_bd = samp_mean + 1.96*std_err)
```

Motivating example—T test iii

```
## # A tibble: 2 x 5
##   feed    samp_mean std_err low_bd upp_bd
##   <fct>       <dbl>   <dbl>  <dbl>  <dbl>
## 1 linseed      219.    15.1   189.   248.
## 2 soybean      246.    14.5   218.   275.
```

- We are interested in whether different feed leads to differences in weight.

Motivating example—T test iv

- One way to formalize this into a hypothesis test is to test whether the *mean weight* is the same for both groups:

$$H_0 : \mu_S = \mu_L.$$

- Our estimators are the sample means for each group.
- In STAT 1150, we saw that we can use the t statistic to perform a t-test for two means.

Motivating example—T test v

```
library(infer)
# By default, it assumes unequal variance
dataset |>
  t_test(formula = weight ~ feed)
```

```
## # A tibble: 1 x 7
## statistic t_df p_value alternative estimate
lower_ci upper_ci
## <dbl> <dbl> <dbl> <chr> <dbl> <dbl> <dbl>
## 1 -1.32 23.6 0.198 two.sided -27.7 -70.8 15.5
```

Motivating example—T test vi

- We get a lot of information out of this:
  - The t statistic is -1.32
  - A 95% confidence interval for the *mean difference* is $(-70.8, 15.5)$.
  - The p-value is 0.198
- Notice that the confidence interval includes the null value, i.e. a difference of zero.
- Overall, we don't have enough evidence to reject the null hypothesis at significance level $\alpha = 0.05$.

Exercise

Using the `gss` dataset in the `infer` package, perform a t-test on the number of hours worked per week (`hours`) and whether a person has a college degree (`college`). Is the difference in average hours significant?

Solution i

```
library(infer)

gss |>
  t_test(formula = hours ~ college)
```

```
## # A tibble: 1 x 7
## statistic t_df p_value alternative estimate
lower_ci upper_ci
## <dbl> <dbl> <dbl> <chr> <dbl> <dbl> <dbl>
## 1 -1.12 366.  0.264 two.sided -1.54 -4.24 1.16
```

Solution ii

- The p-value is 0.264, and therefore we do not reject the null hypothesis of equal average hours worked per week.

Pearson's Chi-Squared Test i

- This is a test for frequency tables (aka contingency tables).

cyl/gear	3	4	5
4	1	8	2
6	2	4	1
8	12	0	2

Pearson's Chi-Squared Test ii

- The main idea is as follows. Let p be the number of cells (e.g. 9 in the example above).
- For each cell, you have the observed count O_i, and an expected count E_i coming from the null hypothesis.
- The test statistic is

$$X^2 = \sum_{i=1}^{p} \frac{(O_i - E_i)^2}{E_i}.$$

- Under the null hypothesis, X^2 approximately follows a chi-square on $(n-1)(m-1)$ degrees of freedom, where n is the number of rows and m, the number of columns.

Example i

- A common null hypothesis used for 2-way contingency tables (like above) is **independence between the two categories**.
- Recall the definition of independence: $P(A, B) = P(A)P(B)$.
- In other words, for two categorical variables X, Y, we have:

$$P(X = x, Y = y) = P(X = x)P(Y = y).$$

Example ii

- Let's add row and column totals to our contingency table:

cyl/gear	3	4	5	Total
4	1	8	2	11
6	2	4	1	7
8	12	0	2	14
Total	15	12	5	32

Example iii

- We have 32 observations; 7 have 6 cylinders, and 12 have 4 gears. Therefore, the probability of having 6 cylinders and 4 gears is equal to

$$\frac{7}{32} \times \frac{12}{32} \approx 0.0820.$$

- Therefore, the *expected* number of observations in the cell is 2.625.

Example iv

- Here's what the *expected* contingency table looks like, under the null hypothesis of independence between the two categorical variables:

cyl/gear	3	4	5
4	5.156	4.125	1.719
6	3.281	2.625	1.094
8	6.562	5.250	2.188

Example v

- Pearson's chi-squared test allows us to measure how *different* this expected table is from the observed table.

```
# Need to transform gear and cyl into factors
mtcars |>
  mutate(gear = factor(gear),
         cyl = factor(cyl)) |>
  chisq_test(formula = gear ~ cyl)
```

Example vi

```
## # A tibble: 1 x 3
##   statistic chisq_df p_value
##       <dbl>    <int>   <dbl>
## 1      18.0        4 0.00121
```

- The p-value is 0.00121. Therefore, we can reject the null hypothesis of independence at significance level $\alpha = 0.05$.
- In other words, we have evidence of **dependence** between the number of gears and the number of cylinders.

Exercise

Using the `gss` dataset in the `infer` package, perform a chi-squared test on the independence between whether a person went to college (`college`) and their opinion of family income (`finrela`).

Solution i

```
gss |>
  chisq_test(college ~ finrela)
```

```
## Warning in stats::chisq.test(table(x), ...): Chi-squa
## incorrect
```

```
## # A tibble: 1 x 3
##   statistic chisq_df   p_value
##       <dbl>    <int>     <dbl>
## 1      30.7        5 0.0000108
```

Remarks

- Pearson's chi-squared test can be used with other types of null distribution.
    - E.g. Hardy-Weinberg equilibrium in genetics; Benford's Law in fraud detection.
- The chi-squared is only an approximation. It can sometimes fail:
    - When you have a small row/column count for a particular category.
    - When you have a small cell count.
- The function `chisq_test` expects a tidy data frame. So you may have to pivot long your tabular data first.

Multiple testing

- Rare events eventually happen if you look (or sample) long enough.
- Even under the null hypothesis, large test statistics can happen.
- This means that if you do 20 hypothesis tests, **even if the null hypothesis is true**, you expect to reject one test.
- Look at this comic: `https://xkcd.com/882/`.
    - Somebody claims jelly beans cause acne.
    - Scientists investigate and find no significant result.
    - They start testing each colour separately, and finally find a colour for which the test is significant.

Bonferroni correction i

- The main issue is that the type I error (i.e. rejecting the null hypothesis when it is true) increases as we perform more test on the same data.
- Let's say we have k (independent) null hypotheses $H_{01}, \ldots, H_{0k}$.

Bonferroni correction ii

$$
\begin{aligned}
P(\text{rejecting one of } H_{0j} | \text{all true}) &= 1 - P(\text{not rejecting any } H_{0j} | \text{all true}) \\
&= 1 - \prod_{j=1}^{k} P(\text{not rejecting } H_{0j} | H_{0j} \text{ true}) \\
&= 1 - \prod_{j=1}^{k} (1 - P(\text{rejecting } H_{0j} | H_{0j} \text{ true})) \\
&= 1 - (1 - \alpha)^k .
\end{aligned}
$$

Bonferroni correction iii

- For different values of k, we get the following table (assuming $\alpha = 0.05$):

# Hypotheses	1	2	5	10	20	100
FWER	0.05	0.0975	0.2262	0.4013	0.6415	0.9941

Bonferroni correction iv

- One potential solution: Bonferroni correction.
  - Replace α with α/k for significance.
- Why does it work? For large k, we have

$$1 - \left(1 - \frac{\alpha}{k}\right)^k \approx 1 - \exp(-\alpha) \approx \alpha.$$

- Failure to account for multiple testing can lead to what is known as **p-hacking**.

Example i

- Going back to the `chickwts` dataset: there are 6 types of feed, meaning we have 15 possible pairwise comparisons.

	horsebean	linseed	meatmeal	soybean	sunflower
casein	7e-07	0.0003	0.0987	0.0035	0.82151177
horsebean	-	0.0069	0.0001	0.0002	0.00000002
linseed	-	-	0.0293	0.1980	0.00002374
meatmeal	-	-	-	0.2252	0.04441462
soybean	-	-	-	-	0.00042876
sunflower	-	-	-	-	-

Example ii

- If we rejected whenever the p-value is less than 0.05, we would reject 11 null hypotheses.
    - But this doesn't take multiple testing into account!
- If instead we rejected whenever the p-value is less than $0.05/15 \approx 0.0033$, then we would reject 7 null hypotheses.
    - casein vs horsebean and linseed
    - horsebean vs meatmeal, soybean, and sunflower
    - sunflower vs linseed and soybean

Permutation tests i

- **Permutation tests** are a large family of resampling methods that can be used to test hypotheses of the form

$$H_0 : F = G,$$

  where F, G are the distribution functions of two different samples.
- You can see this as a *generalization* of the t-test in two ways:
  - We replace equality of means by equality of distributions.
  - We don't assume the data follows a normal distribution.

Permutation tests ii

- It can also be used to test for **independence**:
  - If we have two variables X, Y, with F_X, F_Y the marginal distributions and F_{XY} the joint distribution, independence is equivalent to $F_{XY} = F_X F_Y$.
- The main idea is as follows:
  - Let $X_1, \ldots, X_n \sim F$ and $Y_1, \ldots, Y_m \sim G$.
  - If $H_0 : F = G$ holds, then $X_1, \ldots, X_n, Y_1, \ldots, Y_m \sim F$.
  - Furthermore, **any** permutation of these $n + m$ random variables is also a sample from F!
  - This gives us a way to "generate" data under the null hypothesis.

Permutation tests iii

Algorithm

Let $N = n + m$, and let T be the test statistic on the original sample.

1. Permute the observations to get a sample $Z_1, \ldots, Z_N$.
2. Compute the estimate $T^{(k)} = T(Z_1, \ldots, Z_N)$.
3. Repeat these two steps K times.
4. The permutation p-value is given by

$$\hat{p} = \frac{1 + \sum_{k=1}^{K} I(T^{(k)} \geq T)}{K + 1}.$$

A few observations i

- The procedure is usually considered *approximate*, because we are not using all possible permutations.
    - In practice, $K = 1000$ permutations will give a good approximation.
- Assume that our estimator is the t statistic. To compute the permuted estimate $T^{(k)}$, we compute the sample mean and variance of the first n observations, the sample mean and variance of the remaining m observations, and compute the t statistic.
    - Remember: under the null hypothesis, group membership is meaningless!

A few observations ii

- With permutation tests, the goal is to *break* the association in order to mimic the null hypothesis.
- Permutations = Sampling **without** replacement.

Example i

- You could code this up (exercise!), but the `infer` package does all of this for us.
- Let's go back to testing whether the average number of hours worked per week is different depending on whether someone went to college or not:

```
# calculate the observed statistic
observed_statistic <- gss %>%
  specify(hours ~ college) %>%
  calculate(stat = "t")

observed_statistic
```

Example ii

```
## Response: hours (numeric)
## Explanatory: college (factor)
## # A tibble: 1 x 1
##    stat
##   <dbl>
## 1 -1.12
```

Example iii

```
# generate the null distribution with randomization
null_dist_2_sample <- gss %>%
  specify(hours ~ college) %>%
  hypothesize(null = "independence") %>% # Need null=inc
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "t")
```

Example iv

```
# calculate the p value from the randomization-based nu
# distribution and the observed statistic
null_dist_2_sample %>%
  get_p_value(obs_stat = observed_statistic,
              direction = "two-sided")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.254
```

Other test statistics i

- We already saw above that we can use different test statistics for the same null hypothesis.
- On the other hand, you probably noticed that comparing means is probably not strict enough for $H_0 : F = G$.
    - Distributions can be different but have the same mean.
- One way to more directly compare the full distribution is the *Kolmogorov-Smirnov* test statistic:

$$D = \max_{1 \leq i \leq N} \left| F_n(Z_i) - G_m(Z_i) \right|,$$

where F_n, G_m are the empirical CDFs of $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, respectively.

Other test statistics ii

- The asymptotic distribution of D under the null hypothesis is known, but difficult to compute.
- Permutation tests are a simple alternative, but KS is not currenetly implemented in `infer`.