Statistics for AI and CS

Week 8: Recap

Harmen de Weerd

University of Groningen

Fall 2021

What is it all about

Statistics revolves around quantifying confidence

- What can this sample tell me about the population?
  - How accurate are these claims?
- How much work do I have to do to be 95% sure of my conclusion?

Tools of the trade

There are two main tools used in statistics

- Hypothesis testing
- Confidence intervals

Hypothesis testing

Compare two hypotheses about a population parameter θ

- Assume the truth of the null hypothesis H_0
- Compare the p-value against the significance level α
- Draw a conclusion about the null hypothesis H_0

Hypothesis testing

Compare two hypotheses about a population parameter θ

- Hypotheses are never about sample statistics
  - Population parameters are typically Greek letters $\mu, \sigma, \pi, \beta, \rho, p$
  - Sample statistics are typically Latin letters $\bar{x}, b, s, r$
- Hypotheses are always exact
  - Hypotheses do not have words like significant or approximate

- Assume the truth of the null hypothesis H_0
- Compare the p-value against the significance level α
- Draw a conclusion about the null hypothesis H_0

Hypothesis testing

Compare two hypotheses about a population parameter θ

- Assume the truth of the null hypothesis H_0
  - Null hypothesis H_0 makes an exact claim about the value of θ
    - \+ H_0: The mean heights of mothers and daughters are equal
    - \+ H_0: $\theta = 0$
    - \- H_0: The mean heights of mothers and daughters differ
    - \- H_0: $\theta \neq 0$
  - Alternative hypothesis H_1 is typically the complement of H_0
- Compare the p-value against the significance level α
- Draw a conclusion about the null hypothesis H_0

Hypothesis testing

Compare two hypotheses about a population parameter θ

- Assume the truth of the null hypothesis H_0
- Compare the p-value against the significance level α
  - The significance level α is set before the hypothesis test starts
    - Typically, $\alpha = 0.05$
  - The p-value is the probability of observing your sample statistic, or something more extreme, assuming that H_0 is true
    - The lowest level of significance at which you would reject H_0
- Draw a conclusion about the null hypothesis H_0

Hypothesis testing

Compare two hypotheses about a population parameter θ

- Assume the truth of the null hypothesis H_0
- Compare the p-value against the significance level α
- Draw a conclusion about the null hypothesis H_0
  - Reject H_0 if the p-value is low, or fail to reject H_0 if it is high
    - Never accept H_0
    - Never draw conclusions in terms of H_1
  - Explain the conclusion in terms of θ
    - We reject H_0 and conclude that there the mean heights of mothers and daughters are not equal
    - We fail to reject H_0 and conclude that there is no reason to believe that the mean heights of mothers and daughters are not equal

Statistical errors

Type I error

- Rejecting a null hypothesis that is true
- The Type I error rate equals significance level α by definition
  - If the Type I error rate exceeds α, the test is not appropriate

Type II error

- Failing to reject a null hypothesis that is false
- To calculate the Type II error rate, you need to know the actual value of the population parameter
  - Increasing the sample size decreases Type II error rate
  - Increasing α increases the Type I error rate and decreases the Type II error rate

Confidence intervals

There are two types of confidence intervals

- Confidence interval for population parameters
  - Based on sample statistics and hypothesized distributions
    - This confidence interval is different for each sample
  - With C% confidence, the value of a population parameter is found in a C% confidence interval

- Confidence interval for sample statistics
  - Based on population parameters and known distributions
    - This confidence interval is the same across all samples
  - In C% of the cases, the value of a sample statistic falls within the C% confidence interval

Bootstrapping

A bootstrap sample samples from your dataset

- Approximates the population distribution without assumptions
- Can be used for confidence intervals for population parameters

    - Take a sample with replacement from your dataset of the same size as your dataset
    - Estimate the population parameter
    - Repeat many times
    - Create a confidence interval for the population parameter by discarding the lowest and highest 5% of your estimations

Simulation

Simulation simulates the process of drawing samples

- Makes assumptions about the populations distribution
- Can be used for hypothesis testing
  - Draw a sample from the hypothesized distribution
  - Calculate the test statistic
  - Repeat many times
  - Calculate the p-value by counting how often the simulated value is at least as extreme as the observed value in the sample

Simulation

Simulation simulates the process of drawing samples

- Makes assumptions about the populations distribution
- Can be used for hypothesis testing
  - Draw a sample from the hypothesized distribution
  - Calculate the test statistic
  - Repeat many times
  - Calculate the p-value by counting how often the simulated value is at least as extreme as the observed value in the sample
- Can be used for confidence intervals for sample statistics
  - Draw a sample from the hypothesized distribution
  - Calculate the sample statistic
  - Repeat many times
  - Create a confidence interval for the sample statistic by discarding the lowest and highest 5% of your simulated data

Non-parametric methods

Non-parametric methods make few (or no) assumptions about the population distribution

- Continuous data
  - Sign test
  - Wilcoxon signed rank test
  - Wilcoxon rank sum test
- Categorical data
  - Chi squared test for goodness of fit
  - Chi squared test for independence

Sign test

The **sign test** tests the median: $H_0 : m = m_0$

- Works on continuous data
- Only assumes independent observations
- The sign test has a very high Type II error rate

```
> n = 100
> m0 = 0.5
> binom.test(sum(rnorm(n) < m0), n)

        Exact binomial test

data:  sum(rnorm(n) < m0) and n
number of successes = 69, number of trials = 100, p-value = 0.0001
alternative hypothesis: true probability of success is not equal t
95 percent confidence interval:
 0.5896854 0.7787112
sample estimates:
probability of success
                  0.69
```

Wilcoxon signed rank test

The **Wilcoxon signed rank test** tests the median: $H_0 : m = m_0$

- Works on continuous data
- Assumes independent observations from a symmetric distribution

```
> wilcox.test(rexp(10) - 1)

        Wilcoxon signed rank exact test

data:  rexp(10) - 1
V = 4, p-value = 0.01367
alternative hypothesis: true location is not equal to 0
```

Wilcoxon rank sum test

The **Wilcoxon rank sum test** tests two medians: $H_0 : m_1 = m_2$

- Works on continuous data
- Assumes independent observations from populations with equal distribution shapes

```
> wilcox.test(rexp(10), rexp(15))

        Wilcoxon rank sum exact test

data:  rexp(10) and rexp(15)
W = 99, p-value = 0.1963
alternative hypothesis: true location shift is not equal to 0
```

Chi squared test for goodness of fit

The **Chi squared test for goodness of fit** tests two or more proportions: $H_0 : p_1 = \pi_1, p_2 = \pi_2, \ldots, p_k = \pi_k$

- Works on a single categorical variable
- Null proportions must add up to one: $\sum \pi_i = 1$

```
> chisq.test(table(sample(1:5, 100, replace = TRUE)))

        Chi-squared test for given probabilities

data:  table(sample(1:5, 100, replace = TRUE))
X-squared = 6.4, df = 4, p-value = 0.1712
```

Chi squared test for independence

The **Chi squared test for independence** tests for independence of two variables X and Y: H_0 : X and Y are independent

- Works on paired samples X and Y of categorical data

```
> chisq.test(sample(1:5, 100, replace = TRUE), runif(100) < 0.5)

        Pearson's Chi-squared test

data:  sample(1:5, 100, replace = TRUE) and runif(100) < 0.5
X-squared = 8.2852, df = 4, p-value = 0.08167
```

Central limit theorem

For a sufficiently large sample $X_i, (1 \leq i \leq n)$ that are independent and identically distributed with mean μ and standard deviation σ, the mean $\bar{X}$ is approximately normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$

- For normally distributed data, $n = 2$ is probably enough
- For symmetric distributions, $n = 20$ is probably enough
- For skewed distributions $n = 200$ is probably enough

Proportion test

The **proportion test** tests a proportion: $H_0 : p = p_0$

- Assumes independent observations of a binary variable
- R makes use of a normal approximation based on CLT
  - Given $H_0 : p = p_0$, $\mu = n \cdot p_0$ and $\sigma = \sqrt{\frac{p_0(1-p_0)}{n}}$

```
> prop.test(50,80)

        1-sample proportions test with continuity correction

data:  50 out of 80, null probability 0.5
X-squared = 4.5125, df = 1, p-value = 0.03365
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.5092032 0.7286840
sample estimates:
    p
0.625
```

Proportion test

The **proportion test** tests the equality of two proportions:
$H_0 : p_1 = p_2$

- Assumes independent observations of binary variable

```
> prop.test(c(50,20), c(80, 50))

        2-sample test for equality of proportions
        with continuity correction

data:  c(50, 20) out of c(80, 50)
X-squared = 5.3952, df = 1, p-value = 0.02019
alternative hypothesis: two.sided
95 percent confidence interval:
 0.03643263 0.41356737
sample estimates:
prop 1 prop 2
 0.625  0.400
```

t-test

The t-**test** tests a population mean: $H_0 : \mu = \mu_0$

- Population distribution of the mean is approximately normal
- In general, the population standard deviation is unknown
  - To correct for this, the t-test uses the t distribution

```
> t.test(rnorm(20))

        One Sample t-test

data:  rnorm(20)
t = 0.026165, df = 19, p-value = 0.9794
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.4959853  0.5085430
sample estimates:
  mean of x
0.006278833
```

t-test

The **independent samples t-test** tests the equality of two population means: $H_0 : \mu_1 = \mu_2$

- Population distributions of means are approximately normal
- A **paired samples t-test** is a one-sample t-test

```
> t.test(rnorm(20), rnorm(15))

        Welch Two Sample t-test

data:  rnorm(20) and rnorm(15)
t = -0.22431, df = 32.249, p-value = 0.8239
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.6956803  0.5576213
sample estimates:
   mean of x    mean of y
-0.002683389  0.066346082
```

Methods based on normally distributed data

- (Multiple) linear regression
- ANOVA
- Logistic regression

Assumptions of linear regression

Linear regression assumes a linear relationship between response variable Y and explanatory variables X_i

- Works on continuous response variables Y
- The explanatory variables X_i are linearly independent
  - **Collinearity**: some explanatory variables X_i are correlated
- Residuals are normally distributed
  - Note: not their mean, so CLT does not apply
- The variance of the residuals is constant
  - **Heteroskedasticity**: variance is not constant

Linear regression

```
> summary(lm(len ~ supp + dose, ToothGrowth))

Residuals:
    Min      1Q  Median      3Q     Max
-6.600 -3.700  0.373  2.116  8.800

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.2725     1.2824   7.231 1.31e-09 ***
suppVC       -3.7000     1.0936  -3.383   0.0013 **
dose          9.7636     0.8768  11.135 6.31e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 4.236 on 57 degrees of freedom
Multiple R-squared:  0.7038,    Adjusted R-squared:  0.6934
F-statistic: 67.72 on 2 and 57 DF,  p-value: 8.716e-16
```

For every unit increase in `dose`, `len` increases by 9.79636, assuming that `suppVC` remains constant

university of groningen

Linear regression

```
> summary(lm(len ~ supp + dose, ToothGrowth))

Residuals:
    Min      1Q  Median      3Q     Max
-6.600 -3.700  0.373  2.116  8.800

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.2725     1.2824   7.231 1.31e-09 ***
suppVC       -3.7000     1.0936  -3.383   0.0013 **
dose          9.7636     0.8768  11.135 6.31e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 4.236 on 57 degrees of freedom
Multiple R-squared:  0.7038,    Adjusted R-squared:  0.6934
F-statistic: 67.72 on 2 and 57 DF,  p-value: 8.716e-16
```

Marginal tests test the significance of individual coefficients while assuming other variables are constant $H_0 : \beta_i = 0$

Linear regression

```
> summary(lm(len ~ supp + dose, ToothGrowth))

Residuals:
    Min      1Q  Median      3Q     Max
-6.600 -3.700  0.373  2.116  8.800

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.2725     1.2824   7.231 1.31e-09 ***
suppVC       -3.7000     1.0936  -3.383   0.0013 **
dose          9.7636     0.8768  11.135 6.31e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 4.236 on 57 degrees of freedom
Multiple R-squared:  0.7038,    Adjusted R-squared:  0.6934
F-statistic: 67.72 on 2 and 57 DF,  p-value: 8.716e-16
```

The F test tests the explanatory power of the entire model

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$

Linear regression

```
> summary(lm(len ~ supp + dose, ToothGrowth))

Residuals:
    Min      1Q  Median      3Q     Max
-6.600 -3.700  0.373  2.116  8.800

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.2725     1.2824   7.231 1.31e-09 ***
suppVC       -3.7000     1.0936  -3.383   0.0013 **
dose          9.7636     0.8768  11.135 6.31e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 4.236 on 57 degrees of freedom
Multiple R-squared:  0.7038,    Adjusted R-squared:  0.6934
F-statistic: 67.72 on 2 and 57 DF,  p-value: 8.716e-16
```

Coefficient of determination R^2 is the percentage of variation in the explanatory variable Y that is explained by the model

university of groningen

Model selection

Coefficient of determination R^2 is the percentage of variation in the explanatory variable Y that is explained by the model

- Adding variables to a regression model can not decrease R^2
- To determine whether one model is better, you can use
  - Adjusted R^2: larger is better
  - Akaike Information Criterion AIC: smaller is better
    - `stepAIC` automatically finds minimal AIC
  - Bayesian Information Criterion BIC: smaller is better

ANOVA

ANOVA tests the equality of multiple means:

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$

- Assumes normality of the data and equality of variances
- Equivalent to a linear regression model with only categorical explanatory variables

```
> summary(aov(len ~ supp + dose, data = ToothGrowth))
             Df Sum Sq Mean Sq F value   Pr(>F)
supp          1  205.4   205.4   11.45   0.0013 **
dose          1 2224.3  2224.3  123.99 6.31e-16 ***
Residuals    57 1022.6    17.9
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Two-way ANOVA

Interaction happens when the effect of X_1 on Y depends on the value of X_2

- The effect of dosage on tooth length depends on supplement
- The effect of supplement on tooth length depends on dosage

```
summary(aov(len ~ supp * dose, data = ToothGrowth))
             Df Sum Sq Mean Sq F value   Pr(>F)
supp          1  205.4   205.4  12.317 0.000894 ***
dose          1 2224.3  2224.3 133.415  < 2e-16 ***
supp:dose     1   88.9    88.9   5.333 0.024631 *
Residuals    56  933.6    16.7
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Tukey Honestly Significant Differences

ANOVA can only show *that* there is a difference, not what the difference is

- Tukey HSD performs pairwise *t*-tests with p-values adjusted to account for multiple tests
- Only valid if ANOVA results in a significant effect

```
> TukeyHSD(aov(len ~ factor(dose), data = ToothGrowth))
Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = len ~ factor(dose), data = ToothGrowth)

$`factor(dose)`
        diff       lwr       upr     p adj
1-0.5  9.130  5.901805 12.358195 0.00e+00
2-0.5 15.495 12.266805 18.723195 0.00e+00
2-1    6.365  3.136805  9.593195 4.25e-05
```

Logistic regression

Logistic regression constructs a linear relationship between binary variable Y and explanatory variables X_j

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \sum \beta_j X_j$$

- Uses generalized linear model with a binomial linking function
- Does not assume normality

Logistic regression

Logistic regression constructs a linear relationship between binary variable Y and explanatory variables X_j

- For every unit increase in `site`, the odds of `spam` are *multiplied* by $e^{0.48181}$

```
> summary(glm(spam ~ ., data=emails10, family=binomial))

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.11277    0.03805 -29.247  < 2e-16 ***
site         0.48181    0.05935   8.118 4.74e-16 ***
monday      -1.68278    0.22584  -7.451 9.24e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Null deviance: 6297.6  on 5727  degrees of freedom
Residual deviance: 5818.8  on 5717  degrees of freedom
AIC: 5840.8
```