Introduction to hypothesis testing

Mark Johnson

Macquarie University
Sydney, Australia

February 27, 2017

Outline

Introduction

Hypothesis tests and confidence intervals

Classical hypothesis tests

Regression

Sampling-based hypothesis tests

Conclusion

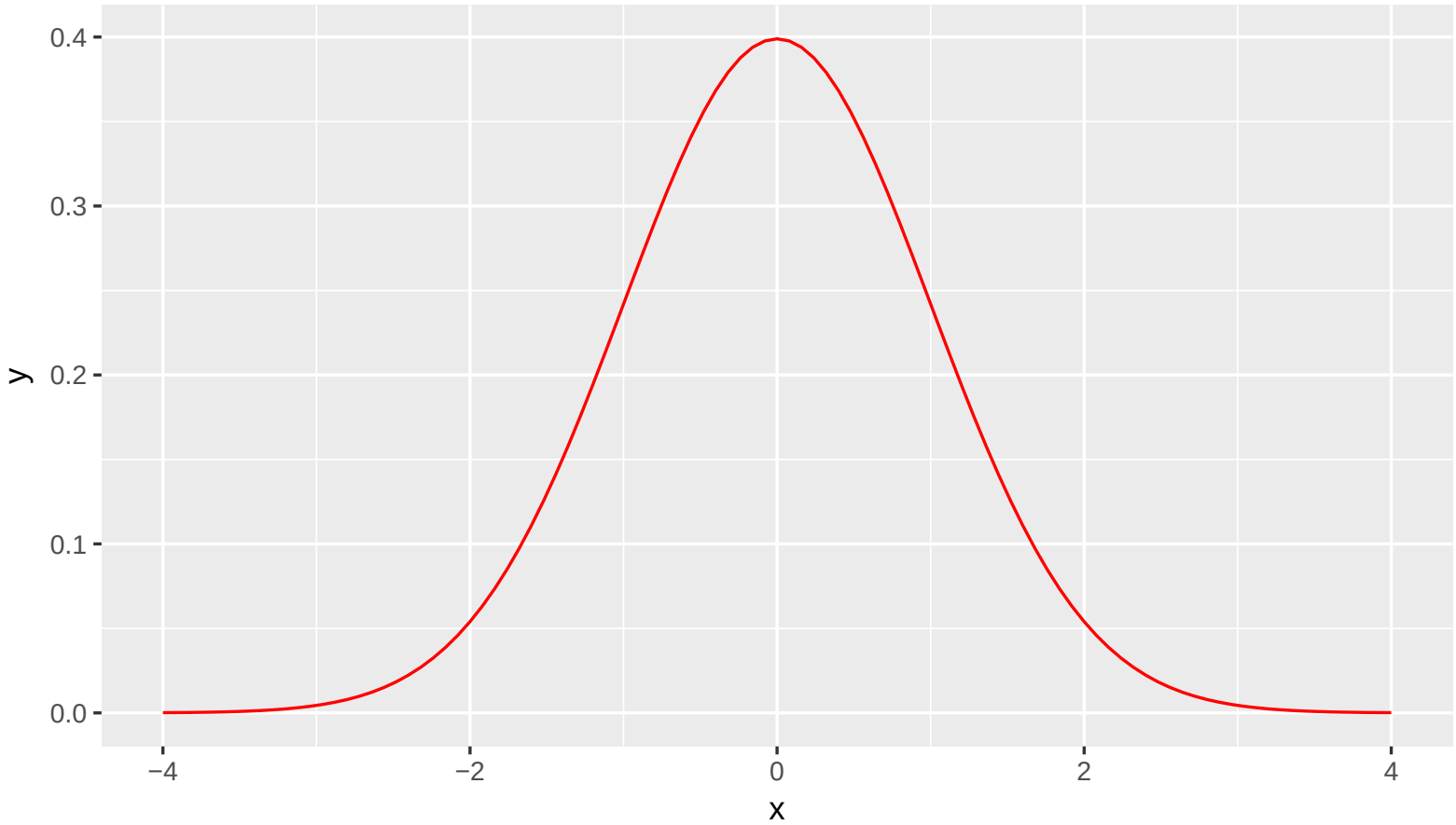
Useful R textbooks from the MQ library web site

- Friedman, Tibshirani and Hastie, 2009, *Elements of statistical learning* (download latest version from authors' web site)
- James, Witten, Hastie and Tibshirani, 2013, *Introduction to statistical learning*: (download latest version from authors' web site)
- Wickham, H. 2009 *ggplot2: Elegant graphics for data analysis*: describes the ggplot2 R graphics package
- Dalgard, P. 2008 *Introductory statistics with R*: general introduction to statistics and R
- Allerhand, M. 2011 *A Tiny handbook of R*: introduces the R programming language

Statistics and Probability

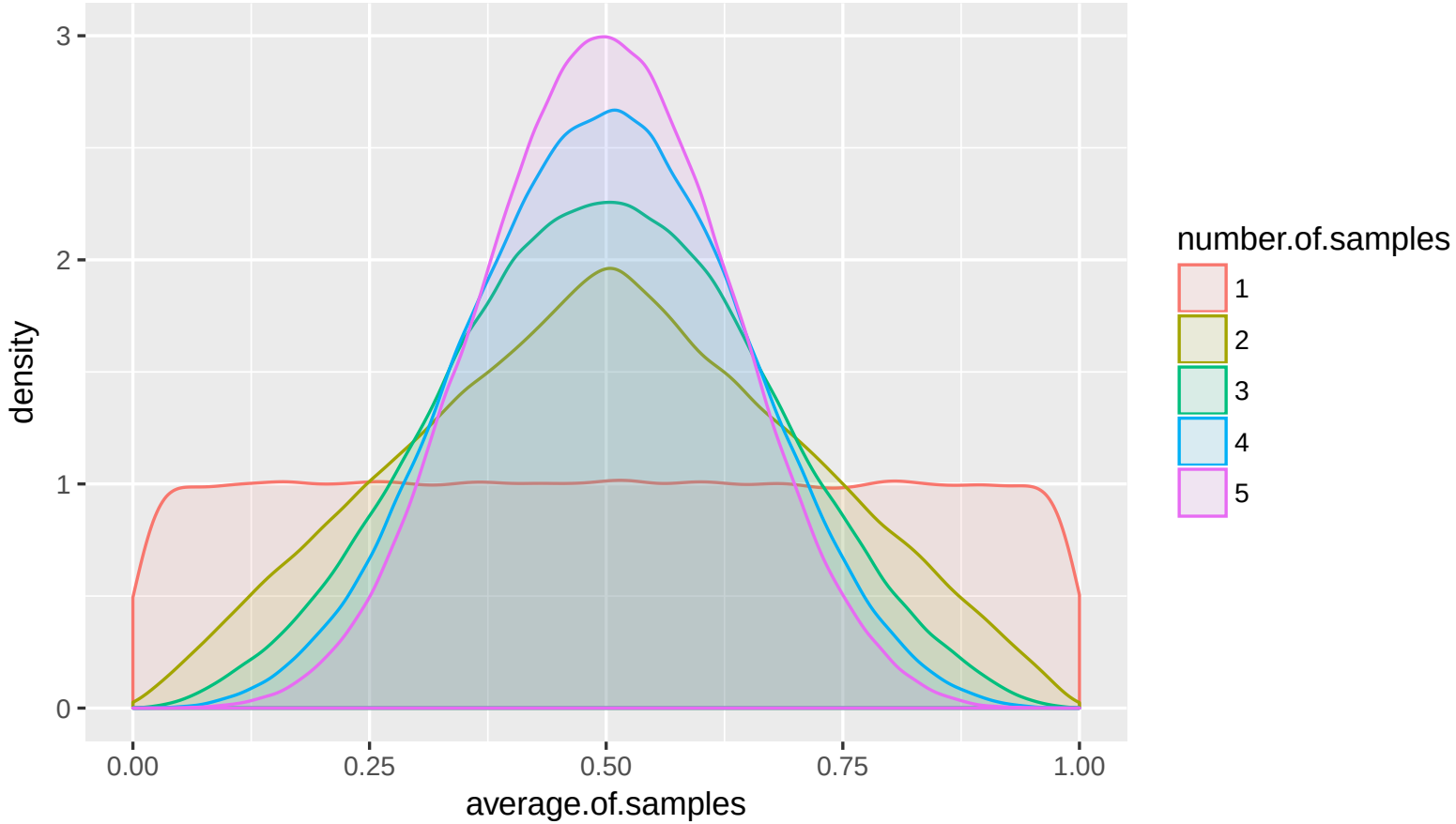
- A statistic is a function of the data (usually chosen to summarise it)
  - example: the *mean* and the *median* are two different statistics
- Probability theory is the mathematics of random phenomena
- Hypothesis tests are statistics that indicate whether a hypothesis is consistent with the data (e.g., “Is this coin fair?”)
- Confidence intervals are statistics that estimate a range of values that contains the true value of a parameter (e.g., “What are the lowest and highest values for the probability of heads?”)
- There’s a general move away from hypothesis tests to confidence intervals

The normal (Gaussian) distribution



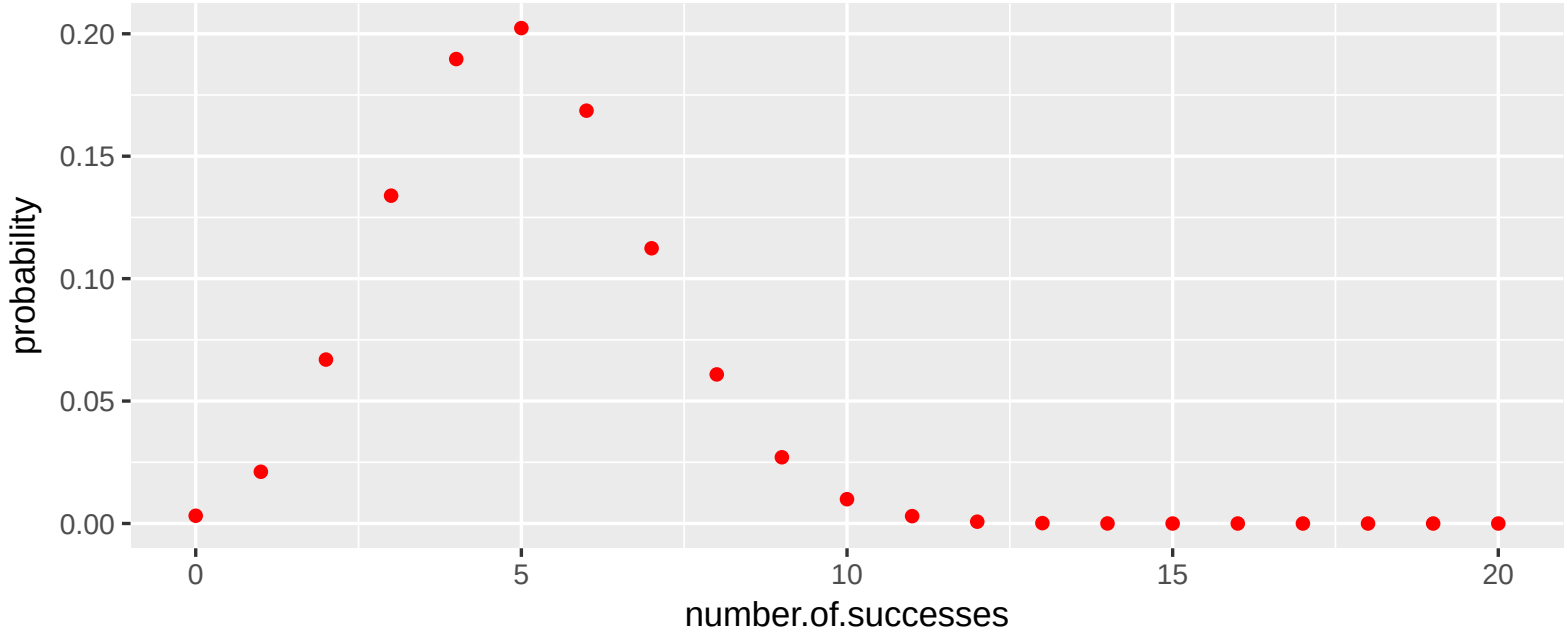
- 0.683 of the probability mass lies in $[-\sigma, \sigma]$
- 0.954 of the probability mass lies in $[-2\sigma, 2\sigma]$
- 0.997 of the probability mass lies in $[-3\sigma, 3\sigma]$

The central limit theorem



- The *central limit theorem* says that *the mean of independent and identically-distributed samples approaches a normal (a.k.a. Gaussian) distribution* as the number of samples grows
  - the normal distribution is usually a fairly good approximation when there are 5 or more samples
  - the *standard deviation of the mean* is approximately $\sigma/\sqrt{n}$, where n is the number of samples

The binomial distribution



- The *binomial distribution* is the distribution of the number of successes in n independent Bernoulli (binary) trials, where each trial has probability p of success
- The binomial distribution has mean $\mu = np$ and standard deviation $\sigma = \sqrt{np(1-p)}$, so

$$\sigma/n = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

Outline

Introduction

Hypothesis tests and confidence intervals

Classical hypothesis tests

Regression

Sampling-based hypothesis tests

Conclusion

Hypothesis testing: motivating examples

- I have a coin, which I'm not sure if is "fair". So I throw it 10 times, and it comes up tails 2 times. Is this evidence that the coin is biased?
- I measure the time it takes for a group of girls to push a button in an experiment, and then I do this for a group of boys. My data show that on average the girls are 10msec faster than the boys. Can I conclude that girls do this task faster than boys, and if so, by how much?
- I've modified my syntactic parser, but I'm not sure if my modifications have really made it more accurate. So I run both the old and the new parsers on the same set of "test sentences" and measure the accuracy of the parses they produce for each sentence. On average my new parser is 2% more accurate than the old one. Is it really better than the old parser, and by how much?

Hypothesis testing vs Predictive modelling

- A *hypothesis test* is intented to determine whether a hypothesis (claim) is true
  - e.g., *coffee causes cancer*,
  - e.g., *algorithm A is faster than algorithm B on a certain kind of data*
  - e.g., *eating more fast food makes you fat*
- A *predictive model* is intended to predict a value as accurately as possible
  - e.g., *predict which individuals are likely to get cancer*
  - e.g., *predict whether algorithm A or algorithm B will run faster on a given data item*
  - e.g., *predict the weight of an individual from the food they eat*

Frequentists and Bayesian approaches

- Frequentist: the probability of an event is the frequency with which it appears in an infinite sequence of replications
- Bayesian: the probability of an event measures the degree of certainty or belief in that event
- Frequentist and Bayesian approaches have different notions of hypothesis testing and confidence intervals
- Frequentist approaches are often more restrictive and unnatural, but computationally simple and better-known in the field
- Bayesian approaches can easily integrate more diverse data, but computationally intensive
- Most “pre-packaged” software implements frequentist approaches, and most examiners/reviewers will expect frequentist analyses, so that’s what we’ll cover here

Is this coin fair?

- Hypothesis H_1: *this coin is not fair*, i.e., $p_{heads} \neq 0.5$
- Null hypothesis H_0: *this coin is fair*, i.e., $p_{heads} = 0.5$
- Data: out of 10 flips, 2 are tails
- Events as or more extreme than the data:
  - 0 tails, 1 tail, 2 tails, 0 heads, 1 head, 2 heads
- Probability of these extreme events under null hypothesis: $p = 0.109$
  - it's conventional to reject the null hypothesis H_0 when p is less than 0.05, 0.01 or 0.001

Hypothesis tests and the null hypothesis

- The Neymann/Pearson/Wald approach to hypothesis testing:
  - given a hypothesis to be tested H_1, formulate an alternative *null hypothesis* H_0
  - pick a *test statistic* T and a *significance level* α
  - calculate the value $T(D)$ of the test statistic on the data D
  - calculate the probability p of data sets with test statistics as or more extreme than $T(D)$
  - if $p < \alpha$ then accept H_1, otherwise reject H_1

Type 1 and type 2 errors

	H_0 is true coin really is fair	H_1 is true coin really is biased
Accept H_0 report coin is fair		Type 2 error false negative
Accept H_1 report coin is biased	Type 1 error false positive	

- In order to bound the probability of Type 2 errors below a small value α, we may have to accept a high probability of making a Type 1 error

What could p_{heads} be?

- Data: out of 10 throws, 8 are heads
- The *maximum likelihood estimate* $\widehat{p}_{heads} = 0.8$, but 8/10 heads is not that unlikely if $p_{heads} = 0.7$
- A *95% confidence interval* is a statistic such were we to flip coins with various values of p_{heads} 10 times, 95% of the time p_{heads} would be within the confidence interval
  - A 95% confidence interval p_{heads} for this data is $[0.444, 0.975]$
- Confidence intervals can be derived from hypothesis tests
  - 0.5 is in the 95% confidence interval for p_{heads} $\iff$ $H_0 : p_{heads} = 0.5$ is not rejected at the 0.05 level

Warning about implicit stopping rules

- If the significance level $\alpha = 0.05$, then the null hypothesis will be rejected about one in every twenty experiments, *even if the null hypothesis is true*

$\Rightarrow$ If you just keep redoing your experiment, *eventually the results will be significant*

  - E.g., if we keep flipping a fair coin, eventually we'll see 10 heads in a row

- Doing this deliberately is scientific fraud, but it's easy to do this accidentally:
  - e.g., keep adjusting your program/experiment until the results are good
  - this is called a *stopping rule*, and significance levels are affected by the stopping rule
- This can be minimised by first selecting the experimental settings on development data, and then performing a single experiment on the test data

Compound hypotheses and Bonferroni correction

- Often we want to test *multiple hypotheses* at once
  - Example: Model A is better than model B and model C
- If we run a large number of hypothesis tests, some will hold "by chance"
- *Bonferroni correction:* To simultaneously test m hypotheses at a significance level α, test each individual hypothesis at the significance level α/m
  - Example: To test that Model A is better than model B and model C, at level $\alpha = 0.01$, run 2 tests (that A is better than B, and that A is better than C) at the $\alpha = 0.005$ level

Outline

Introduction

Hypothesis tests and confidence intervals

Classical hypothesis tests

Regression

Sampling-based hypothesis tests

Conclusion

Classical hypothesis tests

- These are the tests usually found in statistics text books
  - Statistical software packages (like R) provide good implementations of these
- Not computationally intensive (devised before modern computers)
- The *test statistic* is the sum of individual item scores
- The test usually relies on the *Central Limit Theorem* and a *Normal approximation*

Unpaired vs. paired tests

- Some test data is a set of *paired observations*
  - E.g., the predictions of two different classifiers on the *same set of test items* is paired data
  - E.g., the number of people who survive after two different treatments is *not* paired data
- In general it is possible to use an unpaired statistical test on paired data
  - An unpaired test usually has *less power* than a paired test, but the results are still correct
- Using a paired test on unpaired data produces meaningless results

Parametric vs. non-parametric tests

- A *parametric test* assumes that the test statistic is distributed according to some family of distributions (usually the Normal distribution).
  - often reasonable if there is a sufficient number of observations (Central Limit Theorem)
- A *non-parametric test* does not make any assumptions about the distribution of the test statistic.

Two-sample *t*-test

- A two-sample *t*-test tests whether two sequences of real-valued samples come from distributions with different means.
  - this is a parametric test, which assumes that both sequences are normally distributed with the same variance
- Example: *Is the highway miles-per-gallon better in 2008 than in 1999?*

```
t.test(hwy~year, data=mpg)

##
##  Welch Two Sample t-test
##
## data:  hwy by year
## t = -0.032864, df = 231.64, p-value = 0.9738
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.562854  1.511572
## sample estimates:
## mean in group 1999 mean in group 2008
##           23.42735           23.45299
```

See Dalgaard (2008) section 5.3

Two-sample Wilcoxon test

- A two-sample Wilcoxon test tests whether two sequences of real-valued samples come from distributions with different medians
  - it rank orders the values, and tests the distribution of ranks

  ⇒ tied values can be problematic for this test
- It is more robust but less powerful than the two-sample t-test

```
wilcox.test(hwy~year, data=mpg)

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  hwy by year
## W = 6526, p-value = 0.5377
## alternative hypothesis: true location shift is not equal to 0
```

See Dalgaard (2008) section 5.5

Paired t-test

- A paired t-test is used when there are two measurements on each item. The statistics are basically one-sample tests of the difference between the two measurements.
  - paired tests are more powerful than unpaired tests
  - this is a parametric test, which assumes that the differences are normally distributed
- Example: *Is the highway miles-per-gallon better than the city miles-per-gallon?*

```
t.test(mpg$hwy, mpg$cty, paired=TRUE)

##
##  Paired t-test
##
## data:  mpg$hwy and mpg$cty
## t = 44.492, df = 233, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  6.289765 6.872628
## sample estimates:
## mean of the differences
##                6.581197
```

See Dalgaard (2008) section 5.6

The matched-pairs Wilcoxon test

- The matched-pairs Wilcoxon test is a non-parametric version of the paired t-test
- Ties are ignored

```
wilcox.test(mpg$hwy, mpg$cty, paired=TRUE)

##
##  Wilcoxon signed rank test with continuity correction
##
## data:  mpg$hwy and mpg$cty
## V = 27495, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

See Dalgaard (2008) section 5.7

Outline

Introduction

Hypothesis tests and confidence intervals

Classical hypothesis tests

Regression

Sampling-based hypothesis tests

Conclusion

What is linear regression?

- *Regression* estimates the relationship between two or more random variables
- In *simple linear regression* there is a *response or predicted variable* Y and a *explanatory or predictor variable* X, which we assume are related by:

$$Y \sim \alpha + \beta X + N(0, \sigma^2)$$

  where $N(0, \sigma^2)$ is a normal distribution with zero mean and standard deviation σ.
- Given data $D = ((x_1, y_1), \ldots, (x_n, y_n))$ the goal of simple linear regression is to find the *regression coefficient* β and the *intercept* α
  - β is the slope of the line relating X and Y
  - α is the expected value of Y when $X = 0$
- A *Generalised Linear Model* can fit a *Logistic Regression* model to discrete (e.g., binary) data

Regression on highway and city mpg

```
lm(hwy~cty, data=mpg)

##
## Call:
## lm(formula = hwy ~ cty, data = mpg)
##
## Coefficients:
## (Intercept)          cty
##       0.892        1.337
```

- This says:

$$\text{Hwy} \sim 1.337\,\text{Cty} + 0.892 + N(0, \sigma^2)$$

See Dalgaard (2008) section 6.1

Understanding a model formula

- "~" means "distributed as" or "distributed according to"
- So a formula like

$$\text{Hwy} \sim 1.337\,\text{Cty} + 0.892 + N(0, \sigma^2)$$

  can be read as: *to generate a sample value for* Hwy, *sum the following values:*
  - $1.337 \times$ Cty
  - 0.892
  - a sample from $N(0, \sigma^2)$ (a normal distribution with variance σ^2)

Regression parameter estimates

```
m = lm(hwy~cty, data=mpg)
summary(m)

##
## Call:
## lm(formula = hwy ~ cty, data = mpg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3408 -1.2790  0.0214  1.0338  4.0461
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.89204    0.46895   1.902   0.0584 .
## cty          1.33746    0.02697  49.585   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.752 on 232 degrees of freedom
## Multiple R-squared:  0.9138,Adjusted R-squared:  0.9134
## F-statistic:  2459 on 1 and 232 DF,  p-value: < 2.2e-16
```

See Dalgaard (2008) section 6.1

Using regression to identify significant predictors

- Fit a (logistic) regression model to your experimental results
  - Model predicts each test item
  - Regression software estimates significance of each predictor
- More flexible than classical statistical tests
  - Prefer a classical statistical test if one is appropriate

```
m = lm(hwy~year, data=mpg)
summary(m)

##
## Call:
## lm(formula = hwy ~ year, data = mpg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.4530  -5.4530   0.5726   3.5726  20.5726
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.773e+01  1.737e+02   0.102    0.919
## year        2.849e-03  8.669e-02   0.033    0.974
##
## Residual standard error: 5.967 on 232 degrees of freedom
## Multiple R-squared:  4.655e-06,Adjusted R-squared:  -0.004306
```

Outline

Introduction

Hypothesis tests and confidence intervals

Classical hypothesis tests

Regression

Sampling-based hypothesis tests

Conclusion

Sampling-based hypothesis tests

- More flexible than classical hypothesis tests and regression
  - Can use test statistics which aren't sums of individual scores
  - Example: F-score

$$\text{f-score} = \frac{2 \times \text{ \# of correctly proposed items}}{\text{\# of proposed items} + \text{ \# of true items}}$$

- Computationally intensive
  - Requires generating millions of samples
  - Often requires you to write a program
- High-level idea:
  - Sample a large number of variants of test results, modified in a way that should preserve the test statistic *if the null hypothesis is true*
  - Calculate the test statistic on each sample
  - Count the fraction of samples that have a test statistic at least as large as the actual test results

- Smucker, Allan and Carterette (2007) "A Comparison of Statistical Significance Tests for Information Retrieval Evaluation"
- Berg-Kirkpatrick, Burkett and Klein (2012) "An Empirical Investigation of Statistical Significance in NLP"

Permutation tests

- Hypothesis: Model A has a different f-score than model B on test data D
- Null hypothesis: Model A has the same f-score as model B on test data D
  - $\Rightarrow$ Randomly permuting (swapping) the results for model A and model B on any test item $x \in D$ should have no effect on f-score
- The results are a matrix $R = ((A(x_1), B(x_1)), \ldots, (A(x_n), B(x_n)))$
  - For f-score, results for each test item are (# correct, # proposed, # correctly proposed)
- A *permutation* R' of R is produced by *randomly swapping* each row $(A(x_i), B(x_i))$
- Permutation test:
  - Calculate the test results $R = ((A(x_1), B(x_1)), \ldots, (A(x_n), B(x_n)))$
    Calculate f-score difference $\delta = \text{f-score}(R_{:,1}) - \text{f-score}(R_{:,2})$
    Generate n random permutations R' of data set R:
      Calculate the f-score difference $\delta' = \text{f-score}(R'_{:,1}) - \text{f-score}(R'_{:,2})$
    The significance level α is the fraction of samples for which $|\delta'| > |\delta|$

Bootstrap tests using the shift method

- The Bootstrap tests whether model A has a different f-score to model B on test data sets D' from the same distribution as D
- Bootstrap resampling:
  - Draw $|D|$ items *with replacement* from uniform distribution over D
  - In general, a *bootstrap sample* will have repeated items
- Bootstrap samples D' from D in general will not have zero mean f-score difference δ' (why?)
  - The "shift method" shifts the samples so they do have zero mean
- Bootstrap test with the shift method:

  Calculate f-score difference δ on test data D
  Generate n bootstrap samples based on D:
    Calculate the f-score difference δ' for each sample
  The significance level α is fraction of samples for which $|\delta' - \delta| > |\delta|$

Permutation vs. the Bootstrap

Efron and Tibshirani (1998):

> *Permutation methods tend to apply to only a narrow range of problems. However when they apply, as in testing $F = G$ in a two-sample problem, they give gratifyingly exact answers without parametric assumptions.*
>
> *The bootstrap distribution was originally called the "combination distribution." It was designed to extend the virtues of permutation testing to the great majority of statistical problems where there is nothing to permute. When there is something to permute . . . it is a good idea to do so, even if other methods like the bootstrap are also brought to bear.*

Outline

Introduction

Hypothesis tests and confidence intervals

Classical hypothesis tests

Regression

Sampling-based hypothesis tests

Conclusion

Summary and conclusions

- Many reviewers/examiners will expect you to provide statistical significance results
- Classical statistical methods typically require test statistics that are sums of statistics for individual items
- Modern sampling based methods can work with virtually any test statistics