Significance and Hypothesis testing

Martin Popel

ÚFAL (Institute of Formal and Applied Linguistics)
Charles University in Prague

May 13th 2014, Language Data Resources

Motivation

Reporting significance and confidence intervals is ubiquitous in quantitative research.

Goals of this lecture

- Understand the basic principles (and names).
  Understand papers, e.g.
  "significantly better than the baseline"

Motivation

Reporting significance and confidence intervals is ubiquitous in quantitative research.

Goals of this lecture

- Understand the basic principles (and names).
  Understand papers, e.g.
  "significantly better than the baseline"
- Prevent some common pitfalls and fallacies

Motivation

Reporting significance and confidence intervals is ubiquitous in quantitative research.

Goals of this lecture

- Understand the basic principles (and names).
  Understand papers, e.g.
  "significantly better than the baseline"
- Prevent some common pitfalls and fallacies
- Know how to design your own experiments

Motivation

Reporting significance and confidence intervals is ubiquitous in quantitative research.

Goals of this lecture

- Understand the basic principles (and names).
  Understand papers, e.g.
  "significantly better than the baseline"
  Does it mean "much better"?
- Prevent some common pitfalls and fallacies
- Know how to design your own experiments

Motivation

Reporting significance and confidence intervals is ubiquitous in quantitative research.

Goals of this lecture

- Understand the basic principles (and names).
  Understand papers, e.g.
  "significantly better than the baseline"
  Does it mean "much better"? No!
  Don't use "significant" unless you can prove it!
- Prevent some common pitfalls and fallacies
- Know how to design your own experiments

Motivation

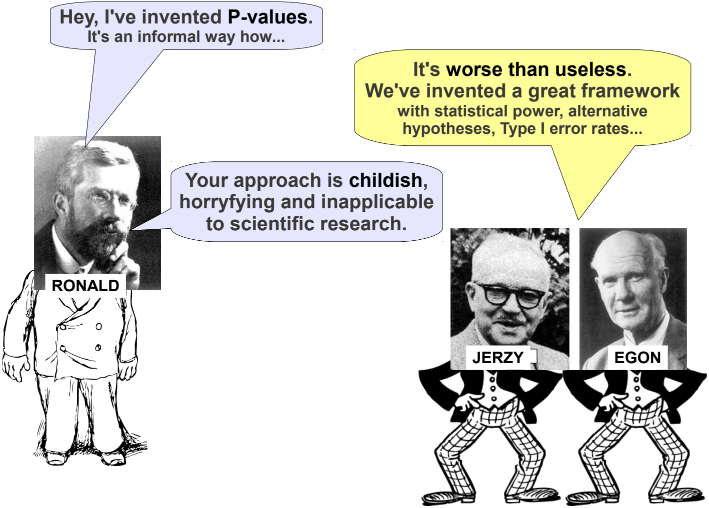
Reporting significance and confidence intervals is ubiquitous in quantitative research.

Goals of this lecture

- Understand the basic principles (and names).
  Understand papers, e.g.
  "significantly better than the baseline ($p < 0.05$)"
  Does it mean "much better"? No!
  Don't use "significant" unless you can prove it!
  So what does it mean?
- Prevent some common pitfalls and fallacies
- Know how to design your own experiments

Fisher vs. Neyman & Pearson

They were rivals, their approaches are not compatible.



Recap: Statistics

What is a statistic?

Recap: Statistics

What is a statistic?

measure (function) of the data, e.g.

- mean ($\bar{X}$, μ),
- standard deviation (s, σ), variance (s^2, σ^2),
- median, Xth quantile,
- for difference tests: difference mean, difference median,...
- BLEU, LAS, F_1-score,...

Recap: Statistics

What is a statistic?

measure (function) of the sample data or whole population, e.g.

- mean ($\bar{X}$, μ),
- standard deviation (s, σ), variance (s^2, σ^2),
- median, Xth quantile,
- for difference tests: difference mean, difference median,...
- BLEU, LAS, F_1-score,...

Recap: Statistics

What is a statistic?

measure (function) of the sample data or whole population, e.g.

- mean ($\bar{X}$, μ),
- standard deviation (s, σ), variance (s^2, σ^2),
- median, Xth quantile,
- for difference tests: difference mean, difference median,...
- BLEU, LAS, F_1-score,...

Recap: Tests

Tests

- one-sample
- two-sample (difference test)
  - unpaired
  - paired

Recap: Tests

Tests

- one-sample
- two-sample (difference test)
  - unpaired
  - paired
    correlated samples have lower variance of the difference mean

P-value

Null hypothesis (H_0):

- no effect, status quo, what could be expected
- defines a distribution

P-value is:

- "the probability of obtaining a test statistic result at least as extreme as the one that was actually observed, assuming that the null hypothesis is true"
- $p = P(data \text{ or more extreme} | H_0)$
- informal measure of evidence against H_0

P-value is not:

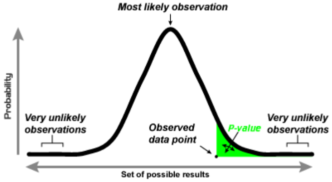
- $P(H_0)$, $P(H_0|data)$, $1 - P(H_A)$ (see Lindley's paradox)
- size or importance of the observed effect
- probability that the measured effect is just a random fluke
- probability of falsely rejecting H_0, i.e. false positive error rate, i.e. Type I error rate

Significance level

Fisher's Significance level:

- popular but arbitrary value is 0.05 (or 0.01 in some areas)
- threshold for p-values (reject H_0 if $p < 0.05$)
- sometimes called α, but should not be confused with Neyman&Pearson's $\alpha =$ Type I error rate.
- should be set before the experiment (prior to data collection)

It is better to report the (rounded) p-value instead of just $p < 0.05$.



A p-value (shaded green area) is the probability of an observed (or more extreme) result arising by chance

Experiment 1: Five heads in a row

- Story: A magician claims to bias a coin toward more heads.
- Experiment: Flip a coin 5 times (i.e. sample size $= 5$).
- Result: HHHHH (i.e. five heads in a row)
- Analysis: p-value $=$
- Conclusion:

Experiment 1: Five heads in a row

- Story: A magician claims to bias a coin toward more heads.
- Experiment: Flip a coin 5 times (i.e. sample size = 5).
- Null hypothesis H_0:
- Result: HHHHH (i.e. five heads in a row)
- Analysis: p-value =
- Conclusion:

Experiment 1: Five heads in a row

- Story: A magician claims to bias a coin toward more heads.
- Experiment: Flip a coin 5 times (i.e. sample size = 5).
- Null hypothesis H_0: $p(head) = p(tail) = 0.5$,
  i.e. the magician has no supernatural abilities, the coin is fair.

- Result: HHHHH (i.e. five heads in a row)
- Analysis: p-value =
- Conclusion:

Experiment 1: Five heads in a row

- Story: A magician claims to bias a coin toward more heads.
- Experiment: Flip a coin 5 times (i.e. sample size = 5).
- Null hypothesis H_0: $p(head) = p(tail) = 0.5$,
  i.e. the magician has no supernatural abilities, the coin is fair.
- Test statistic:
- Result: HHHHH (i.e. five heads in a row)
- Analysis: p-value =
- Conclusion:

Experiment 1: Five heads in a row

- Story: A magician claims to bias a coin toward more heads.
- Experiment: Flip a coin 5 times (i.e. sample size = 5).
- Null hypothesis H_0: $p(head) = p(tail) = 0.5$,
  i.e. the magician has no supernatural abilities, the coin is fair.
- Test statistic: total number of heads

- Result: HHHHH (i.e. test statistic = 5)
- Analysis: p-value =

- Conclusion:

Experiment 1: Five heads in a row

- Story: A magician claims to bias a coin toward more heads.
- Experiment: Flip a coin 5 times (i.e. sample size = 5).
- Null hypothesis H_0: $p(head) = p(tail) = 0.5$,
  i.e. the magician has no supernatural abilities, the coin is fair.
- Test statistic: total number of heads
- Significance level: 0.05 (i.e. confidence level = 95%)

- Result: HHHHH (i.e. test statistic = 5)
- Analysis: p-value =

- Conclusion:

Experiment 1: Five heads in a row

- Story: A magician claims to bias a coin toward more heads.
- Experiment: Flip a coin 5 times (i.e. sample size = 5).
- Null hypothesis H_0: $p(head) = p(tail) = 0.5$,
  i.e. the magician has no supernatural abilities, the coin is fair.
- Test statistic: total number of heads
- Significance level: 0.05 (i.e. confidence level = 95%)

- Result: HHHHH (i.e. test statistic = 5)
- Analysis: p-value $= P(HHHHH \text{ or more} | H_0) = (\frac{1}{2})^5 \doteq 0.03$
  Event HHHHH is significant, p-value < 0.05.
- Conclusion: Reject H_0 (on the 0.05 significance level).
  Either H_0 is false or a highly unprobable event occured.

Experiment 1: Five heads in a row

- Story: A magician claims to bias a coin ~~toward more heads~~.
- Experiment: Flip a coin 5 times (i.e. sample size = 5).
- Null hypothesis H_0: $p(head) = p(tail) = 0.5$,
  i.e. the magician has no supernatural abilities, the coin is fair.
- Test statistic: total number of heads
- Significance level: 0.05 (i.e. confidence level = 95%)

- Result: HHHHH (i.e. test statistic = 5)
- Analysis: p-value =

- Conclusion:

Experiment 1: Five heads in a row

- Story: A magician claims to bias a coin ~~toward more heads~~.
- Experiment: Flip a coin 5 times (i.e. sample size = 5).
- Null hypothesis H_0: $p(head) = p(tail) = 0.5$,
  i.e. the magician has no supernatural abilities, the coin is fair.
- Test statistic: total number of heads
- Significance level: 0.05 (i.e. confidence level = 95%)
- One vs. two tails: two-tailed test
  or alternative hypothesis H_A: $p(head) \neq 0.5$
- Result: HHHHH (i.e. test statistic = 5)
- Analysis: p-value = $P(HHHHH \text{ or more}|H_0) = 2 \cdot (\frac{1}{2})^5 \doteq 0.06$
  Event HHHHH is not significant, p-value $>$ 0.05.
- Conclusion: Cannot reject H_0 (on the 0.05 significance level).

Experiment 1 moral

One tail vs. two tails: It matters.

p-value-two-tailed $= 2\cdot$ p-value-one-tailed (for symmetric H_0)

Which one is more strict?

Experiment 2: Sample size

Test statistic (x): proportion of heads

- HHHHH (5 heads out of 5 flips): $x = 1$
  $p_{\text{two-tailed}} = \frac{1}{16} \doteq 0.06$
- HHHHHHHHHH (10 heads out of 10 flips): $x = 1$
  $p_{\text{two-tailed}} =$
- HHHHHHTHHH (9 heads out of 10 flips): $x = 0.9$
  $p_{\text{two-tailed}} =$

Experiment 2: Sample size

Test statistic (x): proportion of heads

- HHHHH (5 heads out of 5 flips): $x = 1$
  $p_{\text{two-tailed}} = \frac{1}{16} \doteq 0.06$
- HHHHHHHHHH (10 heads out of 10 flips): $x = 1$
  $p_{\text{two-tailed}} = 2 \cdot \frac{1}{2^{10}} = \frac{1}{512} \doteq 0.002$
- HHHHHHTHHH (9 heads out of 10 flips): $x = 0.9$
  $p_{\text{two-tailed}} =$

Experiment 2: Sample size

Test statistic (x): proportion of heads

- HHHHH (5 heads out of 5 flips): $x = 1$
  $p_{\text{two-tailed}} = \frac{1}{16} \doteq 0.06$
- HHHHHHHHHH (10 heads out of 10 flips): $x = 1$
  $p_{\text{two-tailed}} = 2 \cdot \frac{1}{2^{10}} = \frac{1}{512} \doteq 0.002$
- HHHHHHTHHH (9 heads out of 10 flips): $x = 0.9$
  $p_{\text{two-tailed}} = 2 \cdot \frac{1+10}{2^{10}} = \frac{11}{512} \doteq 0.02$

Experiment 2 morals:

- Sample size matters.
- P-value conflates effect size and our confidence.

Experiment 3: Alternating coin flips

Null hypothesis: fair coin

Test statistic: number of heads

- HTHTHTHTHT:
  $p_{\text{two-tailed}} =$

Test statistic (x): number of “alternations” (“HT” or “TH”)

- HTHTHTHTHT:
  $p_{\text{two-tailed}} =$

Experiment 3: Alternating coin flips

Null hypothesis: fair coin

Test statistic: number of heads

- HTHTHTHTHT:
  $p_{\text{two-tailed}} = 1$

Test statistic (x): number of "alternations" ("HT" or "TH")

- HTHTHTHTHT:
  $p_{\text{two-tailed}} = 2 \cdot \frac{1}{2^9} \doteq 0.004$

Experiment 3: Alternating coin flips

Null hypothesis: fair coin

Test statistic: number of heads

- HTHTHTHTHT:
  $p_{\text{two-tailed}} = 1$

Test statistic (x): number of "alternations" ("HT" or "TH")

- HTHTHTHTHT:
  $p_{\text{two-tailed}} = 2 \cdot \frac{1}{2^9} \doteq 0.004$

Experiment 3 morals:

- Test statistic matters.

Confidence Interval

Always report confidence interval for a statistic!
E.g. BLEU=12.1 ([10.6; 12.5])

What influences the size of a confidence interval?

Confidence Interval

Always report confidence interval for a statistic!
E.g. BLEU=12.1 (95% CI [10.6; 12.5])

What influences the size of a confidence interval?

- level of confidence (e.g. 95% confidence interval)

Confidence Interval

Always report confidence interval for a statistic!
E.g. BLEU=12.1 (95% CI [10.6; 12.5])

What influences the size of a confidence interval?

- level of confidence (e.g. 95% confidence interval)
- population variance

Confidence Interval

Always report confidence interval for a statistic!
E.g. BLEU=12.1 (95% CI [10.6; 12.5])

What influences the size of a confidence interval?

- level of confidence (e.g. 95% confidence interval)
- population variance
- sample size

How to compute confidence interval?

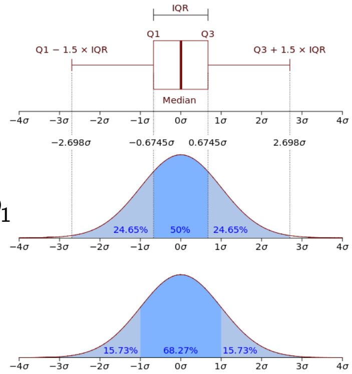
There are three ways

- informal
- traditional normal-based formula
- bootstrapping

How to compute confidence interval?

There are three ways

- informal $Median \pm 1.5 \cdot \frac{IQR}{\sqrt{n}}$
  IQR = Inter-Quartile Range $= Q_3 - Q_1$
  $\sim$ 99% confidence interval
- traditional normal-based formula
- bootstrapping



Normal-based CI

traditional normal-based formula $\bar{x} \pm t \cdot std.err$

- standard error $= \frac{s}{\sqrt{n}} = \frac{\text{sample standard deviation}}{\sqrt{\text{sample size}}}$
  t = t-statistic = function(confidence level, df)
  df = n-1 = degrees of freedom
- ```
  from scipy.stats import t;
  print t.ppf(0.975, 99)
  ```
- Excel, Calc: `TINV(0.05,99)`
- `https://www.wolframalpha.com/input/?i=t-interval`

For example: $n = 100, s = 1, \bar{x} = 10$ the 95% interval is

95% of (population) values lie within this interval. True or false?
  ```

Normal-based CI

traditional normal-based formula $\bar{x} \pm t \cdot std.err$

- standard error $= \frac{s}{\sqrt{n}} = \frac{\text{sample standard deviation}}{\sqrt{\text{sample size}}}$
  t = t-statistic = function(confidence level, df)
  df = n-1 = degrees of freedom
- `from scipy.stats import t;`
  `print t.ppf(0.975, 99)`
- Excel, Calc: `TINV(0.05,99)`
- `https://www.wolframalpha.com/input/?i=t-interval`

For example: $n = 100, s = 1, \bar{x} = 10$ the 95% interval is 10 ± 0.198

95% of (population) values lie within this interval. True or false?

Normal-based CI

traditional normal-based formula $\bar{x} \pm t \cdot std.err$

- standard error $= \frac{s}{\sqrt{n}} = \frac{\text{sample standard deviation}}{\sqrt{\text{sample size}}}$

  t = t-statistic = function(confidence level, df)

  df = n-1 = degrees of freedom

- ```
  from scipy.stats import t;
  print t.ppf(0.975, 99)
  ```
- Excel, Calc: `TINV(0.05,99)`
- `https://www.wolframalpha.com/input/?i=t-interval`

For example: $n = 100, s = 1, \bar{x} = 10$ the 95% interval is 10 ± 0.198

95% of (population) values lie within this interval. True or false? False. We are 95% sure that the population mean lies within this interval.
  ```

Bootstrap

- popular since 90's thanks to faster computers
- distribution-independent
- All the information about the population we have is the sample.
- Resampling produces a similar distribution to repeated sampling from the population.
- The new samples (called "resamples" or "bootstrap samples") must have the same size as the original sample.
- We must sample with replacement. Otherwise all resamples would be identical.
- Sort resamples based on the statistic (mean, BLEU,...).
- Take central 95% of resamples.

Conclusion

Sources and further reading

- http://statslc.com/ youtube videos
- http://en.wikipedia.org/wiki/P-value etc.
- http://vassarstats.net/ can compute test statistic (JS)
- http://www.statisticsdonewrong.com