Lecture 7

Confidence, bootstrapping and mapping disease

Dr. Franco Ubaudi

The Nature of Data
Western Sydney University

Spring 2021

Outline

- Parametric or non-parametric testing
- Population versus sample
- Confidence intervals
- Bootstrapping
- Mapping disease: Kidney Cancer in the US

Introduction

Last time: compared infant birth weights of infants from mothers who smoked to those who did not.

Focus was: compare average birth weight between groups.

Can we compare groups looking at shift in distribution rather than average

Parametric versus non-parametric

- Parametric statistics
  - Uses parameterised statistical distributions
  - e.g. normal distribution $N(\mu, \sigma)$
  - *Makes & depends on assumptions about the data*
- Non-parametric statistics
  - Are distribution-free
  - Or some specified distribution but with unspecified parameters
  - *Makes little or no assumptions about the data*

Wilcoxon-Mann-Whitney test

Suppose we are interested if one group B generally has higher values than the (group A).

We count the number of times an observation in B exceeds one in A.

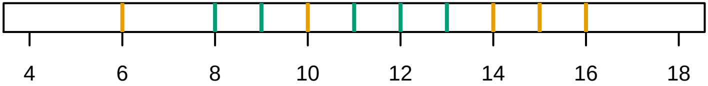


Figure: Location of points from two groups.

- yellow lines = one group, green the other
- For each yellow, count the number of green lines to the left
- Add up the total for all yellow.

Wilcoxon-Mann-Whitney test

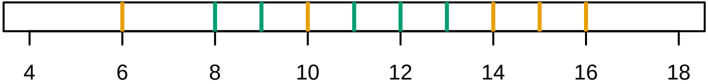


Figure: Location of points from two groups.

0, 2, 5, 5, 5 for a total of 17.

Wilcoxon-Mann-Whitney test

Let n be the size of one group x_i

m be the size of another y_j

Let $U_{ij} = 1$ if $x_i < y_j$ and $U_{ij} = 0$ when $x_i > y_j$

Then

$$U = \sum_j \sum_i U_{ij}$$

Maximum value of U ?

$n \times m$ when all y_j are bigger that all x_i

Minimum possible value?

0

p-values can be computed using tables/approximations provided by R.

This test considers only order

U from Birth Weight

Test if the birth weight is lower when smoking status is “yes”.

bwt	smoke
3429	no
3229	no
3657	yes
3514	no
3086	yes
3886	no

What is U here?

Wilcoxon-Mann-Whitney test

Maternal smoking data $U = 231918$, p-value essentially zero.

For the sales data, it is 1517 and the p-value is 0.064

Hypothesis testing

What do such questions as “Are the means of these two groups equal?” mean?

The mean of the birth weights in smoking/non-smoking in our data, are just numbers.

Why not just see if sample means are equal?

We are trying to make a general inference.

The complete set of all smoking/non-smoking mothers is our population of interest.

The data we have is just a sample

Samples and Populations

- A population contains all individuals or objects of interest.
- Data are collected from a sample, which is a subset of the population.

Hypothesis testing

Suppose the mean birth weight in the population of infants born to non-smoking were μ_1 and μ_2 for smoking.

Group 1 *is non-smoking*
Group 2 *is smoking*

- The emperical mean is an estimate of the population mean ($\bar{x}_1$ is an estimate of μ_1).
- We care if $\mu_1 = \mu_2$
- We try and see if the observed difference in sample means is likely to have occurred if the population means are equal.

Hypothesis testing

Mathematically, we ask if there is any evidence against the null hypothesis

$$H_0 : \mu_1 = \mu_2$$

We have in mind a particular alternative, one of

$$\begin{aligned} H_1 : &\quad \mu_1 \neq \mu_2 \\ H_1 : &\quad \mu_1 > \mu_2 \\ H_1 : &\quad \mu_1 < \mu_2 \end{aligned}$$

We want compute a test statistic and evaluate the chance of seeing something that extreme assuming the null hypothesis.

We have used permutation, Wilcoxon-Mann-Whitney so far (will later do t-tests)

Confidence intervals

Answering a specific question (hypothesis test) may not be enough.

It might tell us there is a difference, but not quantify – the actual reduction in birth weight associated with smoking.

In our data, the difference in sample means $\bar{x}_1 - \bar{x}_2$ is 255.35

This is a guess for the population difference $(\mu_1 - \mu_2)$

But how good?

Confidence intervals

Rather than providing a single point estimate, it would be better if we could get a range estimate on the population difference.

For example, "the true population difference in mean birth weight probably lies between . . . and . . ."

What does probably mean here?

Here, it's about accepting a difference and estimating it.

Confidence intervals

The data is one sample from the population of interest.

Suppose we could take more samples from the population.

Each difference in sample means would be an estimate of the difference in population means.

Some differences would be very large due to random chance.

But most would/should be close to the population mean difference.

Simulation

One group of $n_1 = 50$ observations and a population mean of $\mu_1 = 10$

One of $n_2 = 50$ observations and population mean of $\mu_2 = 15$

There are 1000 sample differences. The true difference is 5

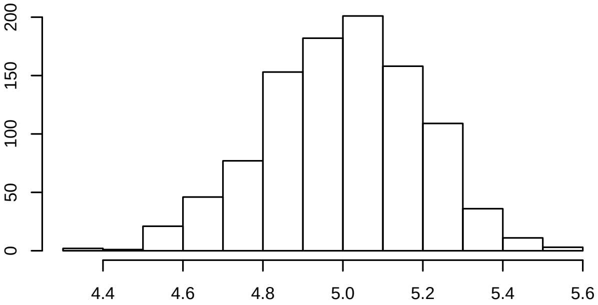


Figure: Difference in means from many samples from the same populations.

Simulation

What we observe:

- Most (95% of them) of the differences lie between 4.61 and 5.36
- The differences are centered on the true (population) difference.
- If we didn't know the true difference, we might say that there is a high chance that the true difference is between 4.61 and 5.36

We actually say that the range 4.61 to 5.36 is a 95% confidence interval for the difference in means.

Bootstrap

Unfortunately, we can almost never do this.

But what we can do, is use multiple samples from the original sample, to predict other samples.

This is called resampling.

Sample with replacement & same sample size $\implies$ bootstrapping

Bootstrap

This is the process of resampling from our data to estimate the distribution of a sample measurement.

We can then compute the difference between the sample means for these resampled data sets, and use it to construct a range or confidence interval.

This whole procedure is called bootstrapping.

Birth weight data

95% of the bootstrap samples lie between 196.44 and 315.16 (the darker lines).

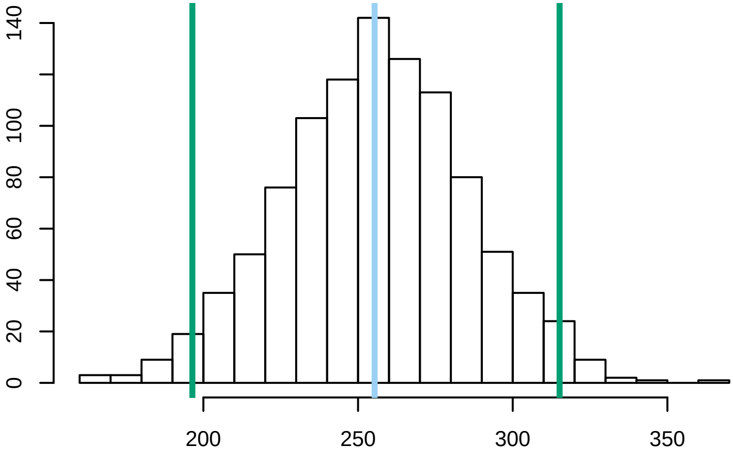


Figure: Boostrap estimate of the distribution of the difference in sample means.

Birth weight data

95% bootstrap confidence interval is between 196.44 and 315.16. The interval boundaries are chosen so that 2.5% of the bootstrap samples are less than 196.44 and 2.5% of the bootstrap samples are greater than 315.16

This is done by taking the 1000 bootstrap sampled differences in means and finding the 25 largest and 25 smallest ($25/1000 = 2.5\%$).

There is nothing special about 95% except that its close to 100%

90% bootstrap confidence interval is between 204.99 and 306.87 (5% above and below).

Sales data

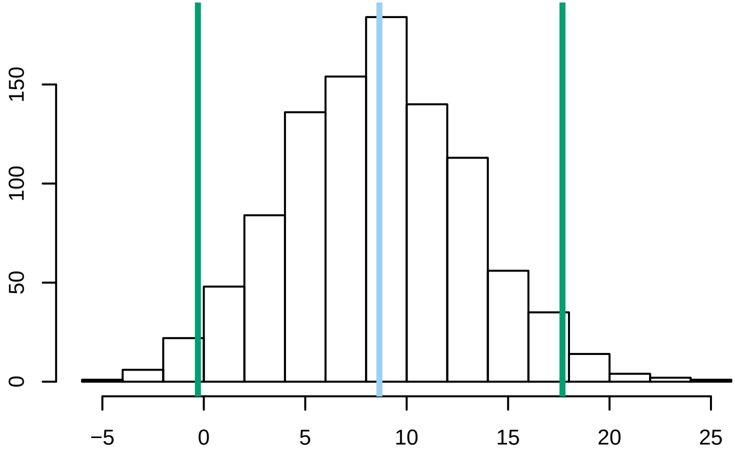


Figure: Bootstrap estimate of the distribution of the difference in sample means.

95% bootstrap confidence interval is between -0.29 and 17.68

Hypothesis tests and confidence intervals are equivalent

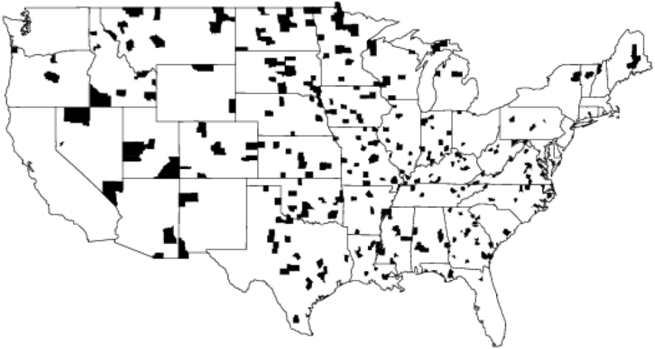
If a 95% confidence interval for the difference in means does not contain the value zero, the the p-value for a two-sided test of whether the difference in means is zero must be greater than 5%

- Confidence interval ≡ CI
- Null hypothesis ≡ H_0
- Assume two-sided hypothesis test
- Critical value ≡ CV

- 95% CI is related to a 5% CV
  (since 100 - 95 = 5)
- If CI contains **zero**, then expect *p-value $>$ CV*
  $\therefore$ don't reject H_0
- Otherwise expect *p-value $<$ CV*
  $\therefore$ reject H_0

Kidney Cancer in the US

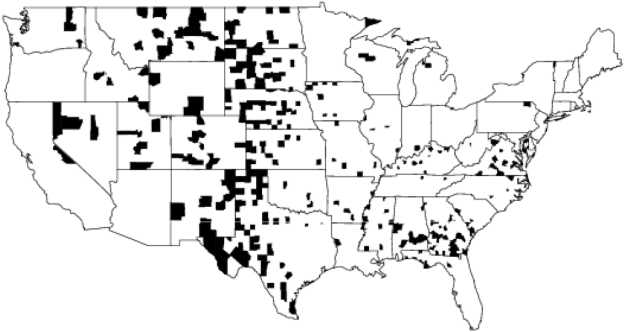
The top 10% of US counties with the highest (adjusted) kidney cancer rates.



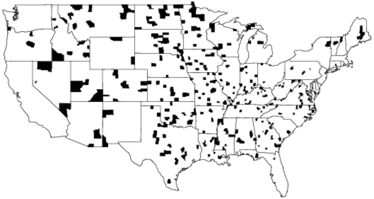
- Why so many in the mid-west?
- One might think it is common in this area.
- This might lead to a search why.

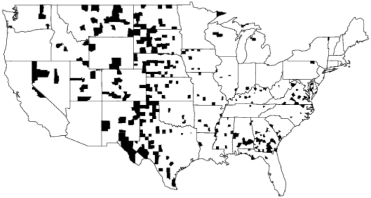
Kidney Cancer in the US

The bottom 10% of US counties with the lowest (adjusted) kidney cancer rates



Kidney Cancer in the US





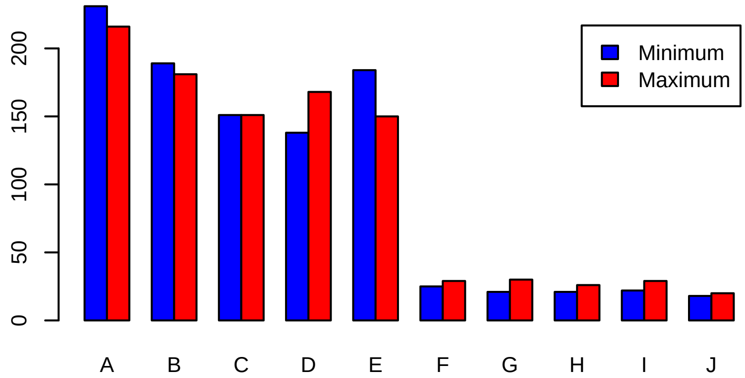
Highest and lowest rates are in very similar areas! What is going on?

Let's simulate.

Simulated disease rates

- Invent cities A–J and simulate people getting a disease with probability 10%.
- A-E have a population 100, F-J have population 1000.
- Generate bar plot of the rate estimated 1000 times.

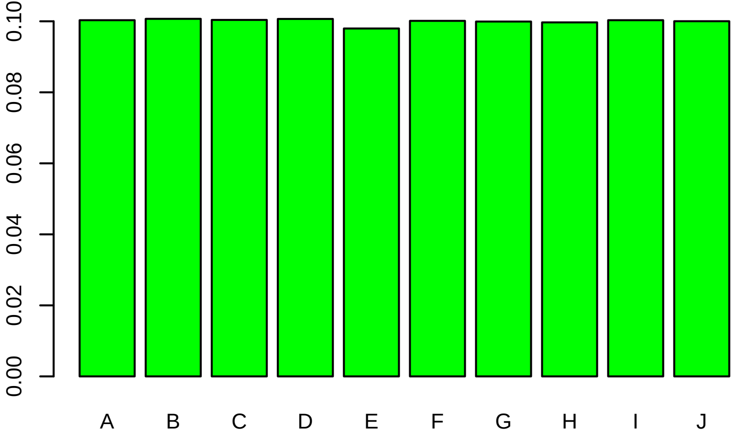
Simulated disease rates



Number of times the minimum (maximum) rate occurs in each "city"

Simulated disease rates

Average rate of disease for 10 cities (1000 simulations).



From counts to rates

The binomial distribution $P(k)$ provides the probability of a certain number k of the population n getting kidney cancer.

Expected number of people who get it is np, variance $np(1-p)$

rate = fraction of people who get cancer.

Expected rate = p, variance in rate = $p(1-p)/n$

$\therefore$ the *larger the population*, the *smaller the variance*.

Back to our simulation

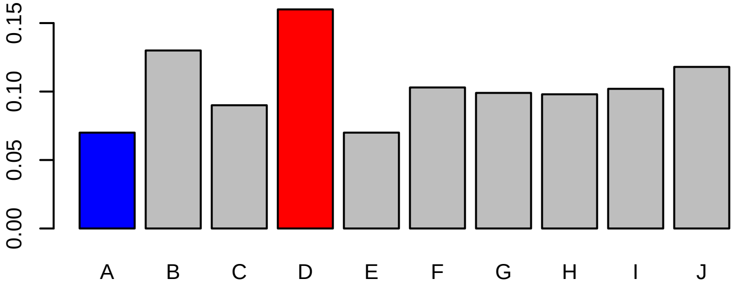
All 10 cities A–J had same underlying rate of disease p.

So why did most of the maxima and minima occur in A–E?

Population sizes.

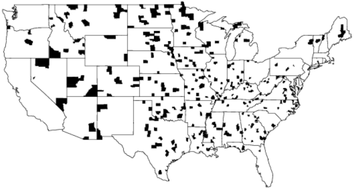
A–E had a population of $n = 100$ and F–J $n = 1000$

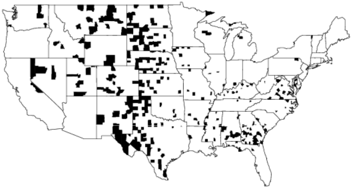
A–E have the same expected rate, but a much higher variability



Kidney Cancer in the US

All cities have the same expected rate, but A–E have much higher variability. So the chances are the largest and smallest values will occur in A–E.





The mid-west counties have the smallest populations.

Moral of the story: be careful how you interpret graphs of rates.

Binomial confidence intervals

Suppose we see 15 seeds germinating out of a plate of 20.
$\therefore$ we estimate the rate of germination as 15/20 or 0.75

If we assume a binomial model, (assuming seeds germinate or not with same rate), then this can be seen as an *estimate* of p

But remember that the 0.75 is the proportion of our sample, not the population. It's a *point estimate*.

How can we compute a confidence interval for the population proportion of seeds that will germinate?

Bootstrap binomial confidence intervals

1. Sample with replacement $n = 20$ seeds from the set with 15 copies of “germinate” and 5 copies of “not germinate”, and count the number of germinates.
2. Repeat 1000 times to obtain a distribution
3. Find interval containing the middle 95%

Bootstrap binomial confidence intervals

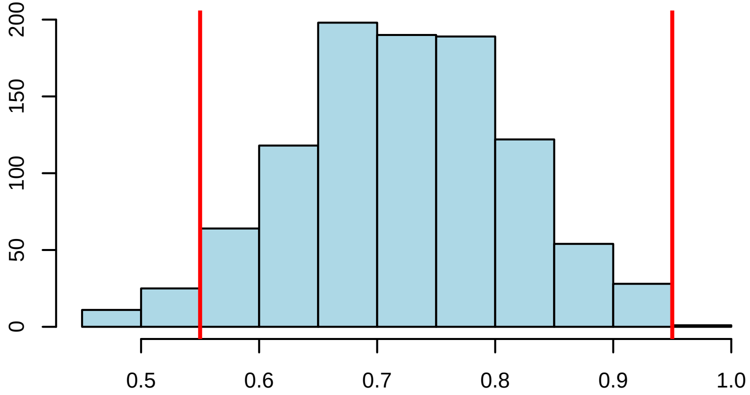


Figure: 95% bootstrap confidence interval for the proportion of germinating seeds.

The bootstrap 95% confidence interval is $[0.55, 0.95]$

Calculate Poisson confidence intervals

Horse kick 200 observations can be bootstrapped

1. Resample with replacement from the 200 observations and compute the mean.
2. Repeat 1000 times
3. Find interval that containing 95% of the examples.

Calculate Poisson confidence intervals

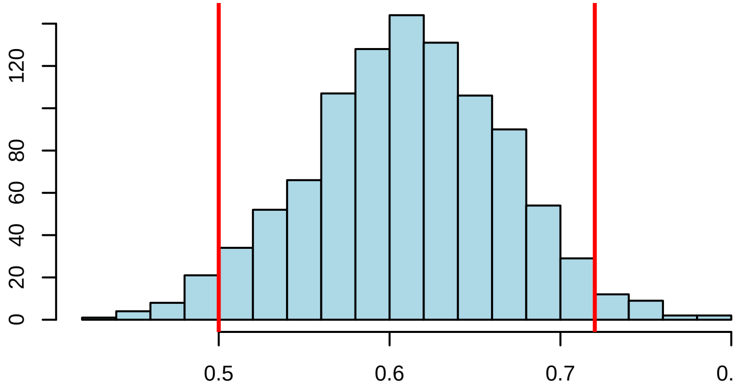


Figure: 95% bootstrap confidence interval for the mean number of deaths by horse kick per year.

95% confidence interval $[0.5, 0.72]$