Tests for qualitative and ranked data

Research Methods for Political Science
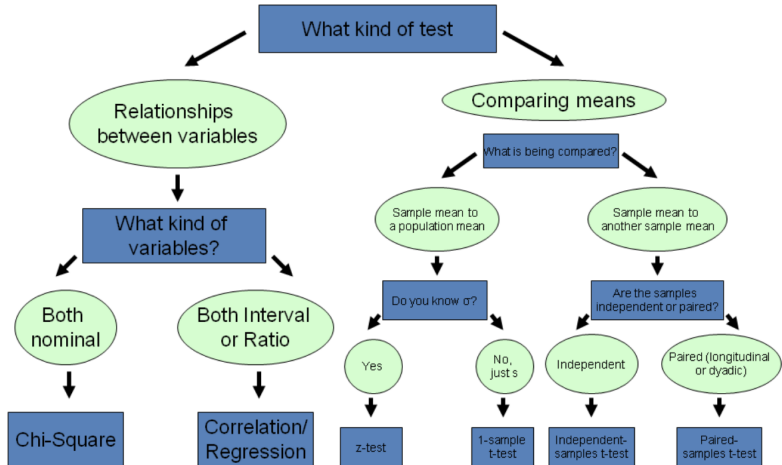
Thomas Chadefaux

Trinity College Dublin

1

Decision Tree

What kind of test

Relationships between variables

Comparing means

What is being compared?

What kind of variables?

Sample mean to a population mean

Sample mean to another sample mean

Both nominal

Both Interval or Ratio

Do you know σ?

Are the samples independent or paired?

Chi-Square

Correlation/ Regression

Yes

No, just s

Independent

Paired (longitudinal or dyadic)

z-test

1-sample t-test

Independent-samples t-test

Paired-samples t-test

Chi-squared test of independence

Chi-squared test of independence: Introduction

- A measure of dependence between two **categorical** variables
- The chi-squared test (also written as χ^2-test) is defined as:

$$\chi^2 = \sum_i \frac{(n_{observed} - n_{expected})^2}{n_{expected}}$$

The degrees of freedom are calculated as $df = (r-1)(c-1)$, where r is the number of rows in the table and c the number of columns.

- If the expected counts and the observed counts are very different, a large value of χ^2 will result. Large values of χ^2 provide evidence against the null hypothesis

χ^2-test: an example

Do dictatorship experience war more often on their territories than democracies?

	Democracy	Dictatorship
No War	40	74
War	3	11

We want to calculate

$$\chi^2 = \sum_i \frac{(n_{observed} - n_{expected})^2}{n_{expected}}$$

But how do we calculate the expected values?

χ^2: calculating expected values for a cross-table

Basic intuition: if the two variables were independent, their relative proportions should be similar to the *marginal* distributions. E.g., the proportion of democracies at war should be similar to the proportion of countries at war. I.e., the probability of a democracy at war if democracy and war were independent would be:

$$\hat{p}_{democracy\&war} = \hat{p}_{democracy} \times \hat{p}_{war}$$

- So first, what is $\hat{p}_{democracy}$? $\hat{p}_{democracy} = 43/128$ and $\hat{p}_{war} = 14/128$. - So if the two variables were independent, we would expect the bottom-left cell to be:
$128 \times \hat{p}_{democracy\&war} = 128 \times \frac{43}{128} \times \frac{14}{128} = 4.7$

χ^2: calculating expected values for a cross-table (cont'd)

- We do the same for every cell and obtain the table of *expected* values:

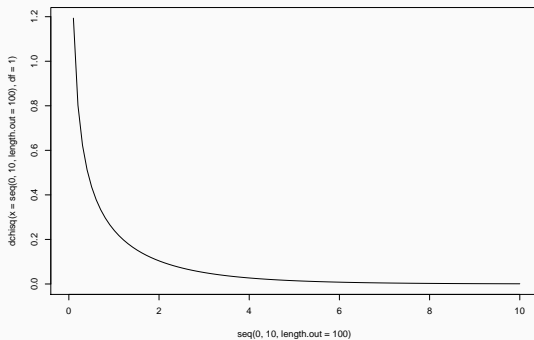|	Democracy	Dictatorship
No War	38.3	75.7
War	4.7	9.3

Now we calculate χ^2 as:

$$\chi^2 = \frac{(40 - 38.3)^2}{38.3} + \frac{(74 - 75.7)^2}{75.7} + \frac{(3 - 4.7)^2}{4.7} + \frac{(11 - 9.3)^2}{9.3} = 1.042$$

- How many df do we have? $(2 - 1) \times (2 - 1) = 1$.

## So what does a $\chi^2(1)$ look like?


```
## [1] 0.3073568
```

So we fail to reject the Null hypothesis that the two variables are independent.

(Aside note on χ^2 distributions)

- Chi square random variables are sums of squared independent standard normal random variables
- So the $\chi^2(1)$ is really a squared standard normal dist.
- Don't take my word for it:

```
hist(rnorm(100000)^2, breaks=100)
```


Histogram of rnorm(1e+05)^2
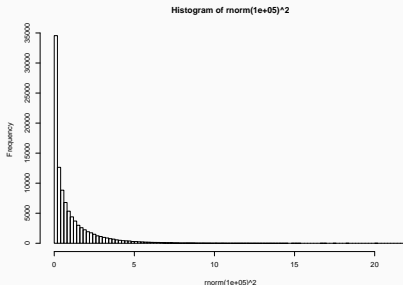
```
hist(rchisq(100000, 1), breaks=100)
```

Histogram of rchisq(1e+05, 1)

χ^2: Another example

Observed:

	Republican	Democrat	Independent
Male	200	150	50
Female	250	300	50

χ^2: Another example (cont'd)

Expected:

	Republican	Democrat	Independent
Male	.4*.45*1000 =180	.4*.45*1000=180	.4*.1*1000=40
Female	.6*.45*1000=270	.6*.45*1000=270	.6*.1*1000=60

so

$$\chi^2 = \frac{(200-180)^2}{180} + \frac{(150-180)^2}{180} + \frac{(50-40)^2}{40} +$$
$$\frac{(250-270)^2}{270} + \frac{(300-270)^2}{270} + \frac{(50-60)^2}{60} = 16.2$$

How many df? $(2-1) \times (3-1) = 2$

```
1-pchisq(16.2,2)
```
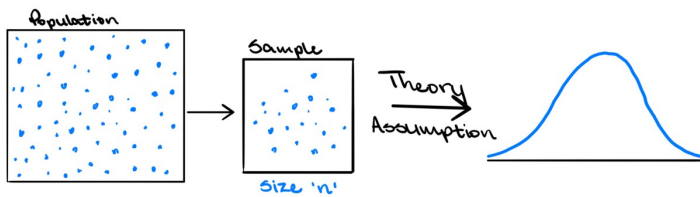
```
## [1] 0.0003035391
```
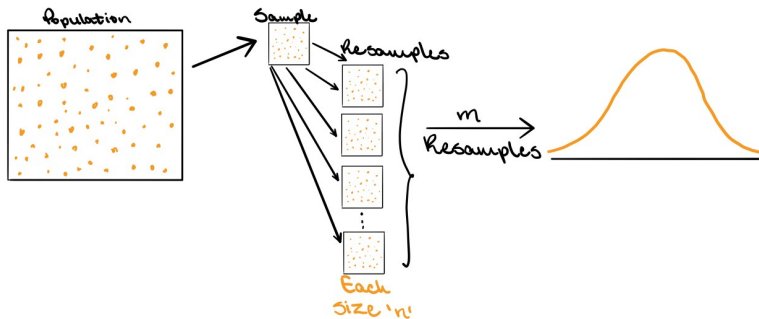
Bootstrapping

The Bootstrap

- Bootstrap: "loop sewn at the top rear or sometimes on each side of a boot to facilitate pulling it on."
- Baron von Muenchhausen pulled himself (and his horse) up by his own bootstraps. . .

Bootstrapping

- Fundamental idea of bootstrapping: use computing power to ask: what would happen if we repeated this method many times?
- The bootstrap is a way of finding the sampling distribution, at least approximately, from just one sample
- Basic idea: Sample from your data, with sample size *n*. Compute your statistic. Repeat many times to obtain the distribution of that statistic.
- From that distribution, you can obtain the mean, SE, CI, etc.

Why bootstrap?

- Issues with normality assumption
- Small sample
- Theoretical distribution of a statistic of interest is complicated or unknown
    - What if you are interested in, say, the *ratio* of means? Or the confidence interval of the median?
    - No simple traditional method for it
- Simpler!
- Appeals directly to the basis of all inference: the sampling distribution that shows what would happen if we took very many samples under the same conditions
- In many fields the preferred way to do inference (sadly less so in the social sciences, but still frequent)

Resampling

1. Resample
   - Typically we only have one sample. The idea here is to create many resamples by sampling *with replacement* from the population
     - 'With replacement' means that we put the observation back before drawing the next. An obs. can therefore be drawn more than once
2. Calculate the statistic for each resample
3. Calculate the statistic of these statistics. E.g.,
   - mean
   - CI

Resampling

The original sample represents the population from which it was drawn. Thus, resamples from this original sample represent what we would get if we took many samples from the population.

The bootstrap distribution of a statistic, based on the resamples, represents the sampling distribution of the statistic

Bootstrapping: Example

- Suppose 4,000 students are supportive of the CETA (Canada-EU trade agreement), whereas 8,000 are opposed
- We want to find the CI of the mean.
- Ideally we use the exact method we learned in the last weeks, but maybe we have a doubt–are the assumptions satisfied, etc.
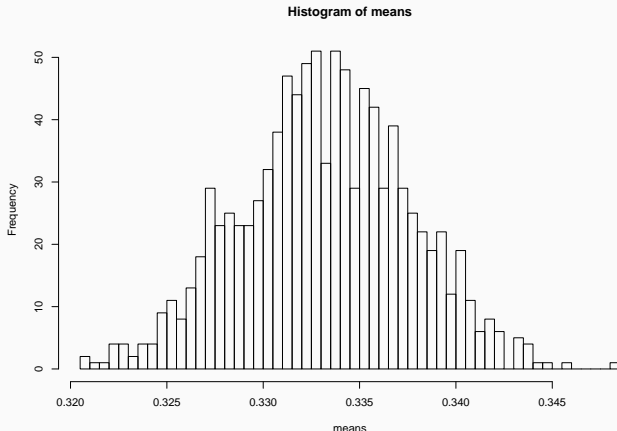- So, bootstrap!

Bootstrapping: Example

```
#Simulate fake data:
CETA <- c( rep(1, 4000), rep(0, 8000))
# bootstrap and calculate means
means <- replicate(n = 1000,
       expr = mean(sample(CETA,
                          size = 12000,
                          replace=T)))
```

`replicate` asks R to repeat a particular expression *n* times (here 1000). So here we are asking R to collect 1000 samples of size 12000, drawn from "population" (not really a population of course) CETA, with replacement.

Bootstrapping: Example

Now we have 1000 samples, and we can see, for example, how they are distributed:

`hist(means, breaks=50)`

Histogram of means

This gives us our bootstrapped sampling distribution! Now we can just pluck values such as the CI:

```
sort(means)[c(25,975)]
```

```
## [1] 0.3246667 0.3416667
```

Not exactly the same as our "proper" result of (0.3249, 0.3418), but pretty close.

Bootstrapping: Example

In fact, increasing the number of resamples gets us very close:

```
means <- replicate(n = 10000,
                   expr = mean(sample(CETA, size = 12000, r
sort(means)[c(250,9750)]

## [1] 0.3248333 0.3419167
```

We can also use this bootstrapped sampling distribution to estimate the standard error:

```
sd(means)
```

[1] 0.004349902

Note that the estimate of the SE is also very close to the value estimated using the formula, i.e., 0.00430

Standard error of the median

Suppose now you wanted to calculate the standard error of the median. It's very simple:

Standard error of the median (ignore this slide and the next!)

Suppose now you wanted to calculate the standard error of the median. It's very simple:

$$f_m(x) = g(c(x))f(x),$$

where

$$g(x) = \frac{(1-x)^{\frac{n-1}{2}} x^{\frac{n-1}{2}}}{B\left(\frac{n+1}{2}, \frac{n+1}{2}\right)},$$

where B is the beta function, $c(x)$ is the cumulative distribution function of the sample distribution, and $f(x)$ is the probability density function of the sample distribution.

Standard error of the median

Now the expected value of the sample median is:

$$\mu_m = \int x f_m(x) dx$$

and the standard deviation of the sample median is:

$$\sigma_m = \sqrt{\int (x - \mu_m)^2 f_m(x) \ dx}$$
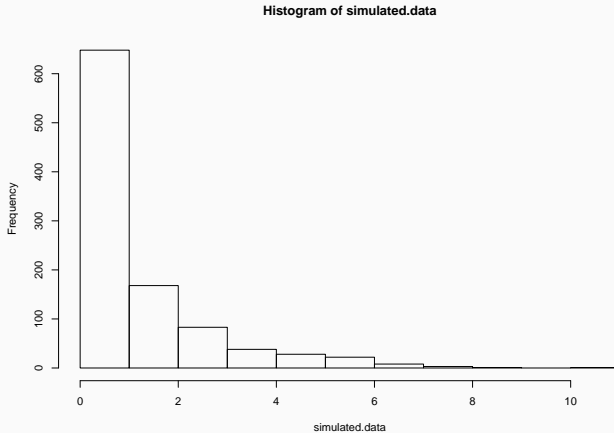
See, simple!

Standard error of the median

Can we do the same thing using bootstrapping? Let's generate random data from, say, a chi-square dist.:

```
n <- 1000
simulated.data <- rchisq(n, df = 1)
```

hist(simulated.data)

Histogram of simulated.data

In-class exercise: calculate the confidence interval of the median

```
medians <- replicate(n = 1000,
                     expr = median(sample(simulated.data,
                                          size = n,
                                          replace=T)))
sort(medians)[c(25,975)]

## [1] 0.4139653 0.5691423
```

and you're done!

Why not ALWAYS use the bootstrap?

If your data has rare extreme values, bootstrapping might never pick these observations, or "under" pick them

$\rightarrow$ It might underestimate the variability in the underlying population.

In practice: why not use both bootstrap **and** standard techniques?

If time allows (unlikely!): Tests for ranked (ordinal) data

Non-parametric tests: introduction

- Methods of inference typically assume normal distribution in the population
- Technically what is assumed is normality of the sample means.
- But when n is small, the Central Limit Theorem does not apply. So we need to assume that the population itself is distributed normally.

- In practice, never perfectly normal, but our methods are robust. They can handle some non-normality
- But sometimes non-normality is a problem. In particular when:
    - Too 'non-normal'
    - Outliers
    - Small sample

What to do when normality is violated?

Basic options:

1. Remove the outliers
   - IFF the outliers are due to, say, bad measurement.
   - But often the outliers are reasonable and valuable info., and you should NOT remove them.
2. Transform your data (e.g., log)
3. Rely on other distributions (e.g., Weibull for survival analysis–not covered here)
4. Bootstrap methods and related (later in this lecture)
5. Nonparametric methods that do not require any specific form for the distribution of the population. Unlike bootstrap and permutation methods, common nonparametric methods do not make use of the actual values of the observations, but instead use counts/ranks of observations.

Comparison of tests based on normal distributions with non-parametric tests

Setting	Normal test	Rank test
One sample	One-sample t-test	Wilcoxon signed rank test
Matched pairs	Paired student t-test	
Two Independent samples	t test	Wilcoxon rank sum test
Several independent samples	F-test	Kruskal-Wallis test.

The Wilcoxon signed-rank test

The Wilcoxon signed-rank test

- Alternative to t-test
- Used to determine whether the **median** of the sample is equal to a known standard value
- Applies to single samples (i.e., difference from zero) or paired samples. e.g.,
    - pre-post test samples. Each individual is measured twice
    - matched samples
- null hypothesis: the numbers of differences in each direction are equal.

Single sample: An example

Suppose we have one sample: $\{-1, 2, 3, -4\}$. The signed rank test goes as follows:

1. sort the numbers: $-4, -1, 2, 3$
2. Assign a rank to the numbers: $1, 2, 3, 4$ (Note: exclude numbers $= 0$)
3. Compute the test statistic:

$$W = \sum_i sign(number_i) \times number_i$$

where $sign(number)$ takes value -1 if the number is negative, 0 if 0, and $+1$ if positive.

Single sample: An example

Here,
$$W = -1 \times 1 + (-1) \times 2 + 1 * 3 + 1 * 4 = 4$$

(Note: R gives a slightly different result because they remove the minimum value. The literature is not unanimous on whether or not it should be removed.)

Two paired samples

Same process, except that instead of raw numbers, we use the difference:

1. Calculate the absolute value of the difference and the sign: $|x_{1,i} - x_{2,i}|$ and $sign(x_{1,i} - x_{2,i})$
2. order and rank the pairs from smallest to largest absolute difference
3. Calculate the test statistic:

$$W = \sum_i sign(x_{2,i} - x_{1,i}) \times rank_i$$

Two paired samples: an example

Suppose we have the following data:

1,2

1,3

4,2

1,5

3, 0

Two paired samples: an example

abs: 1, 2, 2, 4, 3

sign: $+1, +1, -1, +1, -1$

rank: 1, 2.5, 2,5, 4, 3

so W$= 2 + 2.5 - 2.5 + 4 - 3 = 3$

The Wilcoxon rank sum test (aka Mann–Whitney U test)

The Wilcoxon rank sum test

- Tests the null hypothesis that a randomly selected value from one sample is equally likely to be less than or greater than a randomly selected value from a second, independent, sample.
- Alternative to the two sample t-test without assuming normal distributions

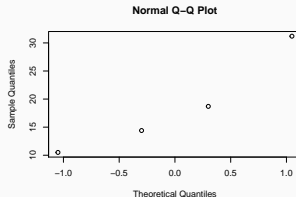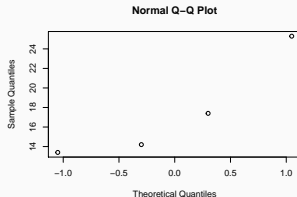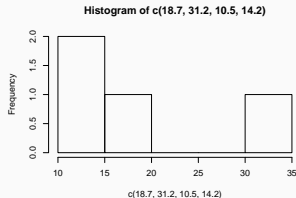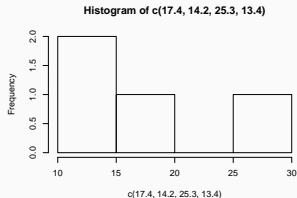## The Wilcoxon rank sum test

Example:

- Suppose we are interested in the effect of TV soap operas on ethnic relations in Burundi.
- We assign a treatment group of 4 individuals to watching TV one hour a day for a month, and a control group of another 4 to watching no TV.
- At the end we ask them to make a donation to a charity that will support the other ethnic group. We record the following donations, in Burundian Francs:

TV?	Donations			
Yes	17.4	14.2	25.3	13.4
No	18.7	31.2	10.5	14.4

Does TV have a significant impact on donations?

First, is the data normally distributed?

Really hard to tell. . . Sample size is too small.

The Wilcoxon rank sum test: Direct Method

For each obs. in one set:

- Count the number of times this first value wins over any obs. in the other set.
- Count 0.5 for ties.
- The sum of wins and ties is U for the first set.

Method 1: Direct Method: an example

Exposure to TV?	Donations			
Yes	17.4	14.2	25.3	13.4
No	18.7	31.2	10.5	14.4

- 17.4 'wins' against 10.5 and against 14.2, so it wins 2 times.
- 14.2 wins against 10.5, so 1 time
- 25.3 wins against 18.7, 10.5 and 14.4, so 3 times
- 13.4 wins against 10.5, so 1 times

So $W = 2 + 1 + 3 + 1 = 7$. The distribution of W is rather complex, so you'll need the software to check its significance.

Method 1: Direct Method: an example (cont'd)

Check with R:

```
A <- c(17.4, 14.2 , 25.3, 13.4)
B <- c(18.7, 31.2, 10.5, 14.4)
wilcox.test(A,B)
```

```
##
##  Wilcoxon rank sum test
##
## data:  A and B
## W = 7, p-value = 0.8857
## alternative hypothesis: true location shift is not equal
```

Recap: the Wilcoxon rank sum test is testing the null hypothesis that $P(X > Y) = 0$.

Consider the following two samples. Calculate W (aka U)

$$A = \{34, 1, 91, 88, 43\}$$

$$B = \{39, 98, 76, 15, 51\}$$

1. Do it manually

Consider the following two samples. Calculate W (aka U)

$$A = \{34, 1, 91, 88, 43\}$$

$$B = \{39, 98, 76, 15, 51\}$$

1. Do it manually

$$U_A = 1 + 0 + 4 + 4 + 2 = 11$$

$$U_B = 2 + 5 + 3 + 1 + 3 = 14$$

2. Write a function in R that will return W for any two samples

```r
myman.u <- function(A, B){
    U <- 0
    for(i in 1:length(A)){
        U <- U + length(which(A[i] > B))
        }
    return(U)
    }
```

Let's test our function

```
# test with new samples
A <- rnorm(100)
B <- rnorm(100)
myman.u(A,B)
```

```
## [1] 5065
```

```
#Verify using R's canned function
myman.u(A,B) == wilcox.test(A,B)$statistic
```

```
##    W
## TRUE
```