Introduction to Handling Data

ECON20222 - Lecture 2 - GUN EXAMPLE version

Ralf Becker and Martyn Andrews

Aim for today

- Explore data
- Review hypothesis testing
- Review simple regression analysis
- Become more familiar with R

Preparing your workfile

We add the basic libraries needed for this week's work:

```
library(tidyverse)      # for almost all data handling tasks
library(readxl)         # to import Excel data
library(ggplot2)        # to produce nice graphics
library(stargazer)      # to produce nice results tables
```

New Dataset - US States Gun Policy

This is the example from the Siegel et al. (2019) paper which attempts to establish whether gun control laws have a causal impact on firearm deaths.

The data are from a variety of sources and some of them are collated in "US_Gun_example.csv". It comprises

- Data for each of the 51 US States and for years 2001 to 2021. This delivers 21*51=1071 observations
- Data on age-adjusted death rates by firearm
- Data on when particular gun laws were in place in different states
- Data for a number of covariates (unemployment rate, number of officers, etc.)

```
merge_data <- read.csv("../data/US_Gun_example.csv") # import
```

This dataset was created with a significant amount of data handling and cleaning.

Gun Law Data

```
str(merge_data)  # prints some basic info on variables
```

```
## 'data.frame':    1071 obs. of  19 variables:
##  $ X                 : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Year              : int  2001 2001 2001 2001 2001 2001 2001 2001 2001 2001 ...
##  $ State_code        : chr  "AK" "AL" "AR" "AZ" ...
##  $ State             : chr  "Alaska" "Alabama" "Arkansas" "Arizona" ...
##  $ Population        : int  633687 4467634 2691571 5273477 34479458 4425687 34328
##  $ Mechanism         : chr  "Firearm" "Firearm" "Firearm" "Firearm" ...
##  $ pol_ubc           : int  0 0 0 0 1 0 1 NA 0 0 ...
##  $ pol_vmisd         : int  0 0 0 0 1 0 1 NA 1 0 ...
##  $ pol_mayissue      : int  0 1 0 0 1 1 1 NA 1 0 ...
##  $ Age.Adjusted.Rate: num  14.83 16.41 15.27 15.92 9.32 ...
##  $ logy              : num  2.7 2.8 2.73 2.77 2.23 ...
##  $ ur                : num  6.28 5.18 4.76 4.72 5.47 ...
##  $ law.officers      : int  1821 15303 7538 18548 106244 15172 9825 4716 2964 630
##  $ law.officers.pc   : num  287 343 280 352 308 ...
##  $ vcrime            : int  3696 19203 12042 28275 210661 15334 11387 4845 4845 1
##  $ vcrime.pc         : num  583 430 447 536 611 ...
##  $ alcc.pc           : num  2.67 1.86 1.74 2.5 2.2 ...
##  $ incarc            : int  3033 24741 11489 27710 157142 14888 17507 NA 6841 724
##  $ incarc.pc         : num  479 554 427 525 456 ...
```

Data Description

Some of the names in `merge_data` are obvious, but let us introduce a few in more detail.

- `Age.Adjusted.rate` - Deaths by firearm for every 100,000 population in a State-Year. Here is a note on age adjustment.
- `ur` - Average annual unemployment rate in a State-Year
- `law.officers.pc` - number of law officers per 100,000 population in a State-Year
- `vcrime.pc` - number of violent crimise (excl. homicides) per 100,000 population in a State-Year

These are the key variables we concentrate on in this data introduction.

```
summary(merge_data[c("Age.Adjusted.Rate","ur","law.officers.pc","vcrime.pc")])
```

```
##  Age.Adjusted.Rate       ur         law.officers.pc    vcrime.pc
##  Min.   : 2.14     Min.   : 2.100   Min.   : 37.08   Min.   :  76.94
##  1st Qu.: 8.84     1st Qu.: 4.204   1st Qu.:255.98   1st Qu.: 268.22
##  Median :11.72     Median : 5.283   Median :296.35   Median : 358.09
##  Mean   :12.05     Mean   : 5.648   Mean   :310.21   Mean   : 393.14
##  3rd Qu.:15.05     3rd Qu.: 6.737   3rd Qu.:343.69   3rd Qu.: 493.70
##  Max.   :33.82     Max.   :13.733   Max.   :894.46   Max.   :1056.47
##                                     NA's   :23
```

Data - State-Years

To find the states and years in the sample:

```
unique(merge_data$State)   # unique finds all the different values in a variable
```

```
##  [1] "Alaska"               "Alabama"              "Arkansas"
##  [4] "Arizona"              "California"           "Colorado"
##  [7] "Connecticut"          "District of Columbia" "Delaware"
## [10] "Florida"              "Georgia"              "Hawaii"
## [13] "Iowa"                 "Idaho"                "Illinois"
## [16] "Indiana"              "Kansas"               "Kentucky"
## [19] "Louisiana"            "Massachusetts"        "Maryland"
## [22] "Maine"                "Michigan"             "Minnesota"
## [25] "Missouri"             "Mississippi"          "Montana"
## [28] "North Carolina"       "North Dakota"         "Nebraska"
## [31] "New Hampshire"        "New Jersey"           "New Mexico"
## [34] "Nevada"               "New York"             "Ohio"
## [37] "Oklahoma"             "Oregon"               "Pennsylvania"
## [40] "Rhode Island"         "South Carolina"       "South Dakota"
## [43] "Tennessee"            "Texas"                "Utah"
## [46] "Virginia"             "Vermont"              "Washington"
## [49] "Wisconsin"            "West Virginia"        "Wyoming"
```

```
unique(merge_data$Year)
```

```
##  [1] 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
## [16] 2016 2017 2018 2019 2020 2021
```

Data - State-Years

To find out how many observations we have for each state (`State`) and also calculate the mean of the above four variables. Use piping technique of the `tidyverse`

```
sel_states <- c("New York", "California","West Virginia","Nebraska")
table1 <- merge_data %>% filter(State %in% sel_states) %>% # only looks at selected
      group_by(State) %>% # groups by State
      summarise(n = n(),
                avg.fds = mean( Age.Adjusted.Rate),
                avg.ur = mean( ur),
                avg.off = mean( law.officers.pc),
                avg.vcr = mean(vcrime.pc)) %>%                  # calculating no of obs
      print()
```

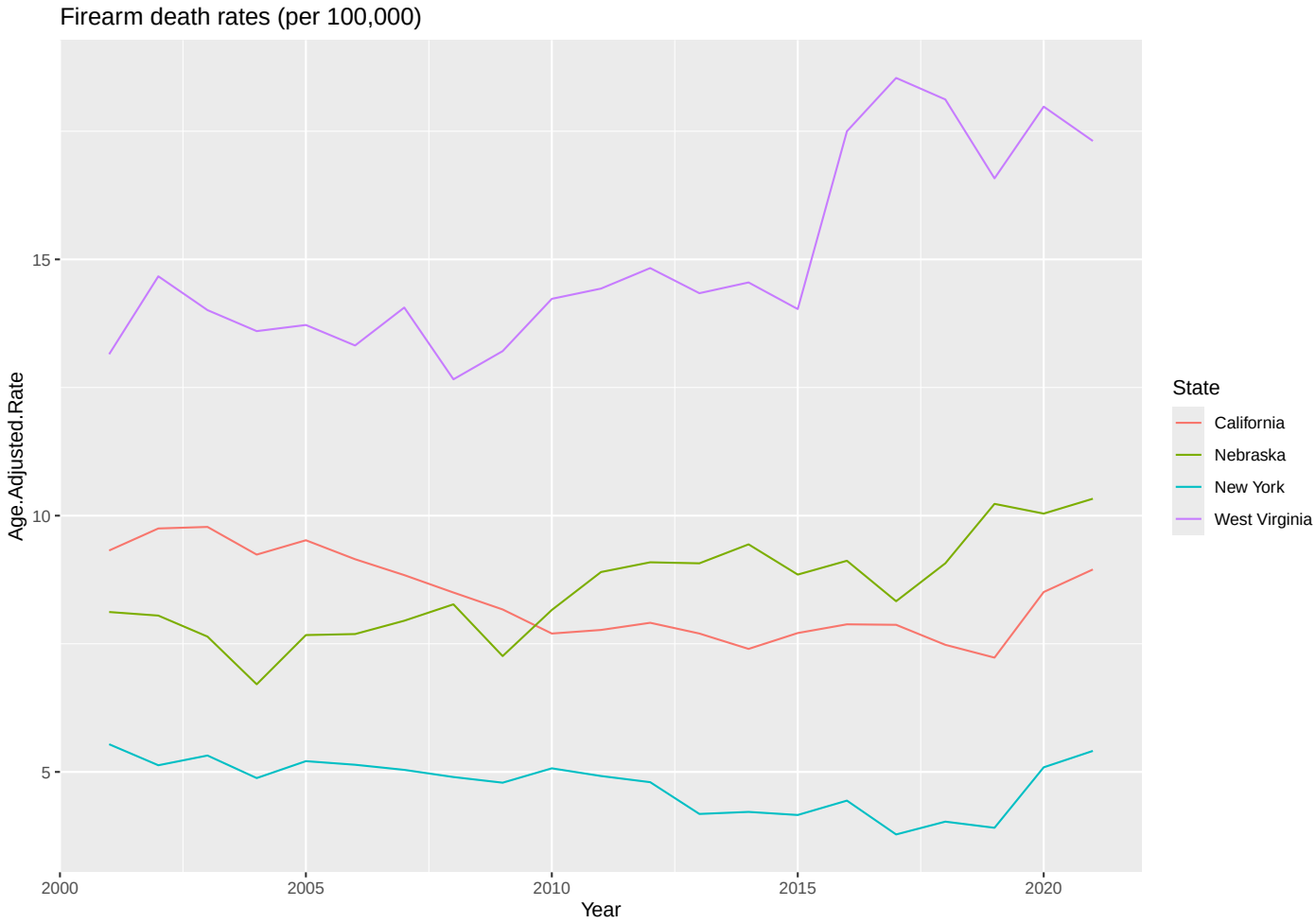
```
## # A tibble: 4 x 6
##   State             n avg.fds avg.ur avg.off avg.vcr
##   <chr>         <int>   <dbl>  <dbl>   <dbl>   <dbl>
## 1 California       21    8.40   7.32    314.    481.
## 2 Nebraska         21    8.57   3.54     NA     297.
## 3 New York         21    4.76   6.21    424.    404.
## 4 West Virginia    21   15.0    6.24     NA     301.
```

For each state ($j = 1, ..., 51$) we have observations from 21 years ($t = 1, ..., 21$). We index observations with the subscript jt. This is what we call a panel.

Data - Some graphical representation

Plotting the number of officers against firearm death rates

Data - Some graphical representation



Data - Some graphical representation

There seems to be a negative relationship between Firearm death rates (`Age.Adjusted.Rate`) and number of officers (`Law.Officers.pc`). **This is variation across states and years.**

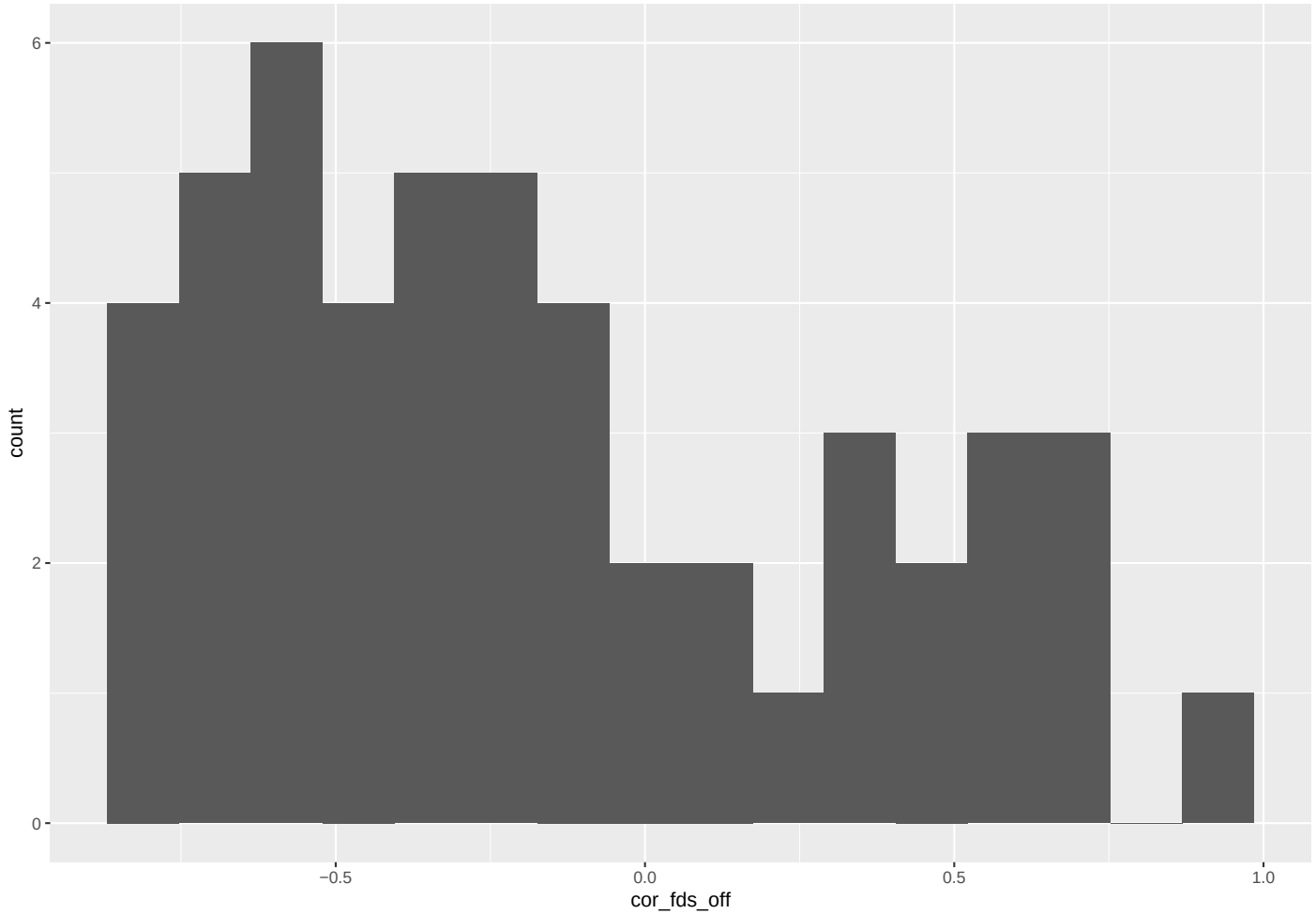
Is there such a relationship inside states as well? **Time/year variation only.**

We calculate correlations for each state:

e.g. Arizona: $Corr_{AZ}(fds_{AZ,t}, off_{AZ,t})$

```
## # A tibble: 8 x 2
##   State       cor_fds_off
##   <chr>             <dbl>
## 1 Alabama          -0.120
## 2 Alaska           -0.623
## 3 Arizona          -0.265
## 4 Arkansas          0.596
## 5 California        0.403
## 6 Colorado         -0.646
## 7 Connecticut      -0.119
## 8 Delaware         -0.320
```

Data - Some graphical representation



Data on Maps

Geographical relationships are sometimes best illustrated with maps.

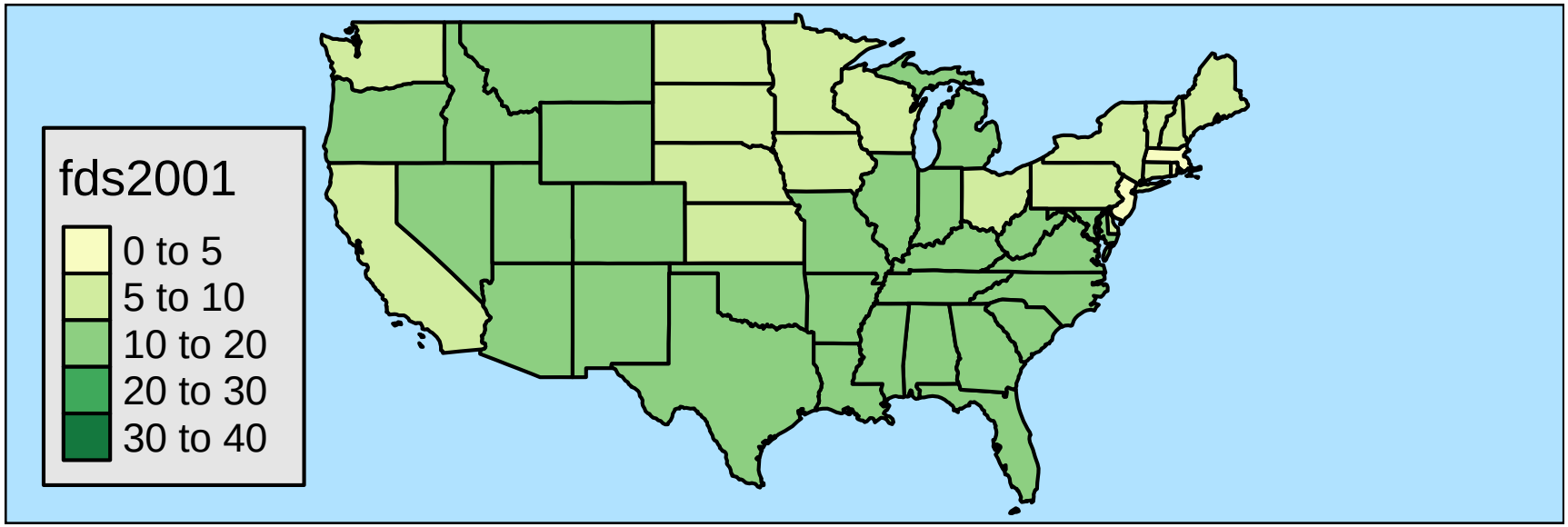
Sometimes these will reveal a pattern.

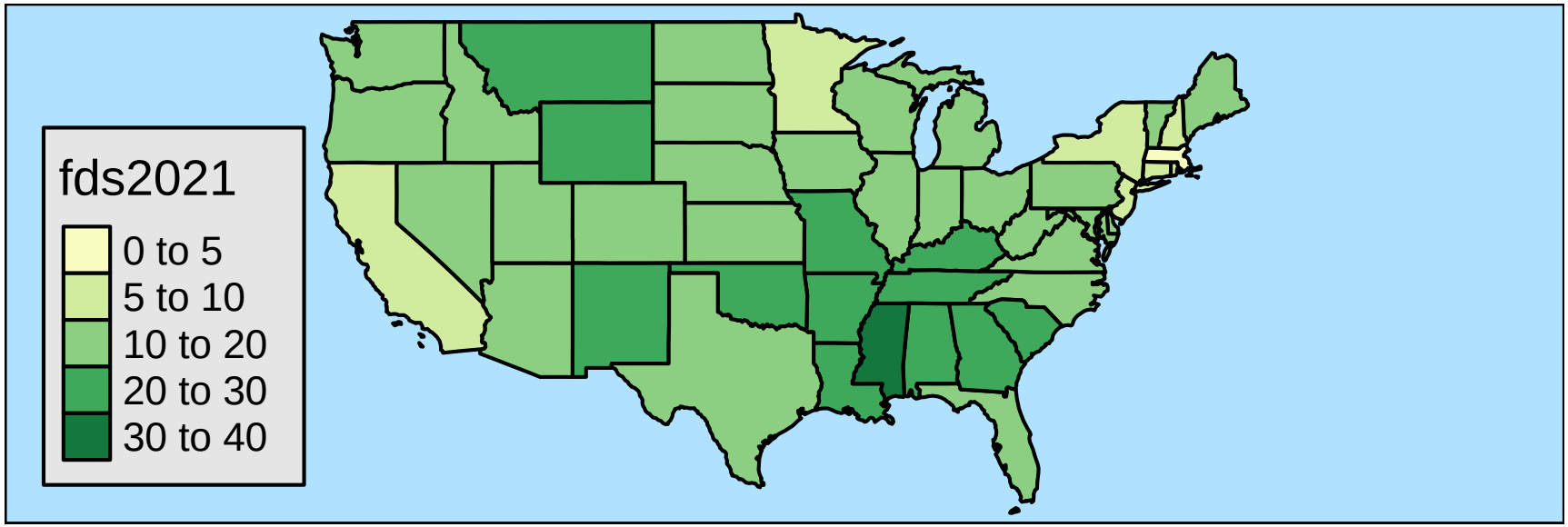
R can create great maps (but it requires a bit of setup - see the additional file on BB). You need the following

- A shape file for each country
- The statistics for each country
- a procedure to merge these bits of information in one data-frame (`merge`)

Let's look at the firearm death rates and how these vary by state.

Data Using Maps: 2001 v 2021





Hypothesis Testing - Introduction

Hypothesis testing is a core technique used in empirical analysis. Use sample data to infer something about the population mean (or correlation, or variance, etc). Hence *inference*.

It is crucial to understand that the particular sample we have is one of many different possible samples. Whatever conclusion we arrive at is not a conclusion with certainty.

Hypothesis Testing - Introduction

Example

Are the average firearm death rates in 2001 and 2021 different from each other?

$$\begin{aligned} H_0 &: \mu_{fds,2001} = \mu_{fds,2021} \\ H_A &: \mu_{fds,2001} \neq \mu_{fds,2021} \end{aligned}$$

When performing a test we need to calibrate some level of uncertainty. We typically fix the Probability with which we reject a correct null hypothesis (Type I error). This is also called the significance level.

Hypothesis Testing - Introduction

Depending on the type of hypothesis there will be a test statistic which will be used to come to a decision.

Assuming that H_0 is true this test statistic has a random distribution (frequently t, N, χ^2 or F). We can then use this distribution to evaluate how likely it would have been to get the type of sample we have if the null hypothesis was true ( ) or obtain .

Decision Rule 1

If that probability, **p-value** < **significance level**, then we reject H_0.
If, however, that **p-value** > **significance level** then we will not reject H_0

Decision Rule 2

If the absolute value of the **test statistic** > the **critical value** (obtain from the Null distribution - see next slide), then we reject H_0.
If, however, the absolute value of the **test statistic** is < the **critical value**, then we will not reject H_0.

Hypothesis Testing - Introduction

Example The test statistic for testing the equality of the average violent crime (vc) in 2001 and 2021.

$$t = \frac{\bar{vc}_{2021} - \bar{vc}_{2001}}{\sqrt{\frac{s_{vc,2021}}{n} + \frac{s_{vc,2001}}{n}}}$$

How is this test statistic distributed (assuming H_0 is true)? **If**

1. The two samples are independent
2. The random variables fds_{2021} and fds_{2001} are either normally distributed or we have sufficiently large samples
3. The variances in the two samples are identical

then $t \sim$

The above assumptions are crucial (and they differ from test to test). If they are not met then the resulting p-value (or critical values) are not correct.

Hypothesis Testing - Example 1

Let's create a sample statistic:

```
test_data_2001 <- merge_data %>%
  filter(Year == 2001)    # pick 2001
mean_2001 <- mean(test_data_2001$vcrime.pc)

test_data_2021 <- merge_data %>%
  filter(Year == 2021)       # pick 2021
mean_2021 <- mean(test_data_2021$vcrime.pc)

sample_diff <- mean_2021 - mean_2001
paste("mean_2021 - mean_2001 =", round(mean_2021,2),
      " - ", round(mean_2001,2), " = ", round(sample_diff,2))
```

```
## [1] "mean_2021 - mean_2001 = 370.23  -  424.34  =  -54.11"
```

Is this different significant?

The difference, 54, is about 13% of the 2001 mean.

Hypothesis Testing - Example 1

Formulate a null hypothesis: ***the difference in population means (mu) in vcrime.pc is equal to 0***. We use the `t.test` function. We deliver the `vcrime.pc` series for both years to `t.test`.

```
t.test(test_data_2021$vcrime.pc,test_data_2001$vcrime.pc, mu=0)  # testing that mu =
```

```
##
##  Welch Two Sample t-test
##
## data:  test_data_2021$vcrime.pc and test_data_2001$vcrime.pc
## t = -1.5375, df = 94.887, p-value = 0.1275
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -123.98112   15.75956
## sample estimates:
## mean of x mean of y
##  370.2264  424.3372
```

The p-value then tells us how likely it is to get a result like the one we got (a difference of -54 or larger) if the null hypothesis was true (i.e. the true population means were the same).

The p-value is 0.1275 and hence it is possible but not very likely that this difference would have arisen by chance if the null hypothesis WAS correct.

Hypothesis Testing - Example 2

What about the difference between 2011 and 2021 though?

```
test_data_2011 <- merge_data %>%
  filter(Year == 2011)
t.test(test_data_2021$vcrime.pc,test_data_2011$vcrime.pc, mu=0)  # testing that mu =
```

```
##
##  Welch Two Sample t-test
##
## data:  test_data_2021$vcrime.pc and test_data_2011$vcrime.pc
## t = 0.37061, df = 99.381, p-value = 0.7117
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -47.90612  69.91315
## sample estimates:
## mean of x mean of y
##  370.2264  359.2229
```

The p-value is 0.7117 and hence

Hypothesis Testing - To reject or to not reject

When comparing between 2011 and 2021 the p-value was 0.71:

When comparing between 2001 and 2021 the p-value was app. 0.13:

- Conventional significance levels are 10%, 5%, 1% or 0.1%
- But what do they mean?

To illustrate we add a random variable (**rvar**) to all observations. The value comes from the identical distribution for all observations, the standard normal ($N(0,1)$ or **rnorm** in R):

```
merge_data$rvar <- rnorm(nrow(merge_data))    # add random variable

test_data <- merge_data
years <- unique(merge_data$Year)  # List of all years
n_years <- length(years)
```

By construction we know that the true underlying mean is identical in all countries.

But what happens if we calculate sample means of **rvar** in all years and then compare between years?

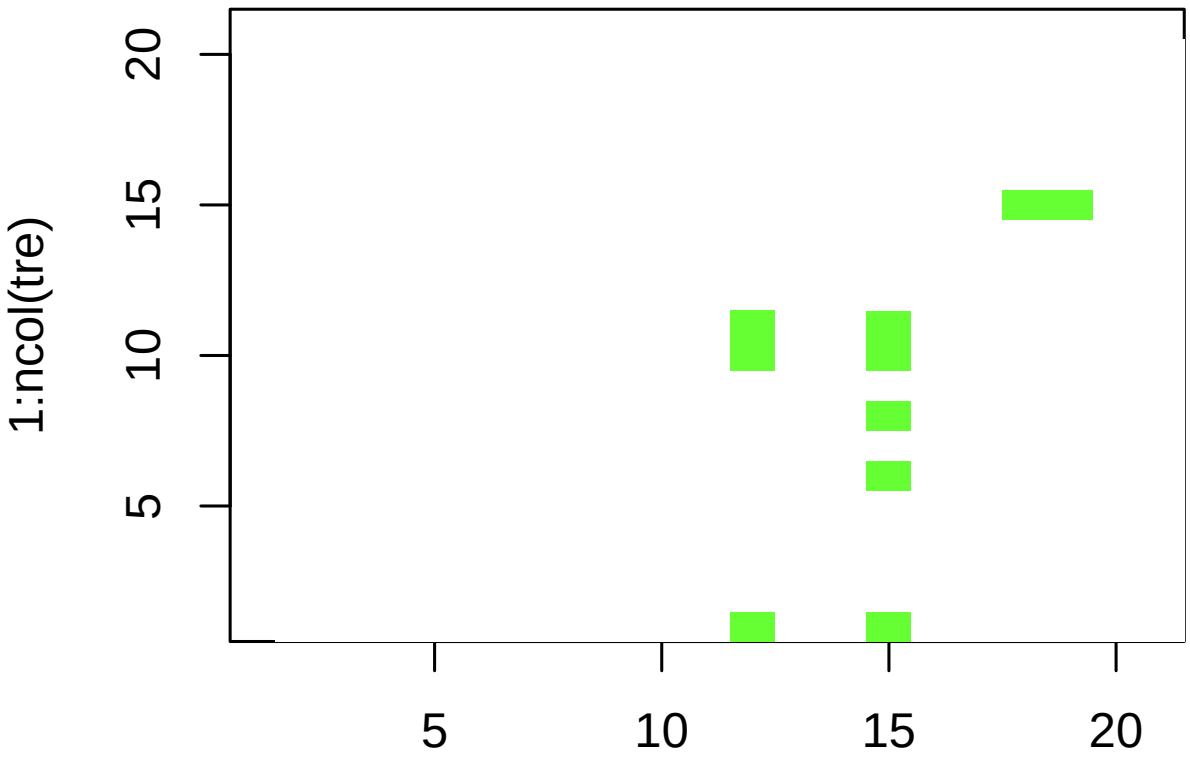
Hypothesis Testing - To reject or to not reject

We have 210 hypothesis tests, all testing a correct null hypothesis. If a p-value is smaller than 10% we decide to reject H_0.

All null hypotheses we know to be true (population means are identical). Let's see how many of these hypothesis tests delivered p-values which are smaller than 10%.

```
## tre
## FALSE  TRUE
##   200    10
```

Let's present a graphical representation of these results. Every green square representing a rejection of the null hypothesis.



Regression Analysis - Introduction

Tool on which most of the work in this unit is based

- Allows to quantify relationships between 2 or more variables
- It can be used to implement hypothesis tests
- However it does not necessarily deliver causal relationships!

It is very easy to compute for everyone! Results will often have to be interpretated very carefully.

Your skill will be to interpret carefully and correctly!!!!

Regression Analysis - Example 1

Let's start by creating a new dataset which only contains the 2001 data.

```
test_data <- merge_data %>%
  filter(Year == 2001)
```

Now we run a regression of the violent crimes per 100,000 population variable (`vcrime.pc`) against a constant only.

$vcrime.pc_i = \alpha + u_i$

```
mod1 <- lm(vcrime.pc~1,data=test_data)
```

Regression Analysis - Example 1

We use the **stargazer** function to display regression results

```
stargazer(mod1, type="text")
```

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                vcrime.pc
## ------------------------------------------------
## Constant                      424.337***
##                                (27.624)
##
## ------------------------------------------------
## Observations                     51
## R2                              0.000
## Adjusted R2                     0.000
## Residual Std. Error       197.274 (df = 50)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

The estimate for the constant, $\widehat{\alpha}$, is the sample mean.

Regression Analysis - Example 1

Testing $H_0 : \mu_{vcrime.pc} = 0$ can be achieved by

```
##
##  One Sample t-test
##
## data:  test_data$vcrime.pc
## t = 15.361, df = 50, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  368.8529 479.8216
## sample estimates:
## mean of x
##  424.3372
```

We can use the above regression to achieve the same:

$t - test = \widehat{\alpha}/se_{\widehat{\alpha}}$

Regression Analysis - Example 2

New dataset which contains the 2001 and 2021 data. Create a dummy variable (1 = if obs from 2021, 0 otherwise)

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                vcrime.pc
## ------------------------------------------------
## Year2021                       -54.111
##                                (35.194)
##
## Constant                      424.337***
##                                (24.886)
##
## ------------------------------------------------
## Observations                     102
## R2                              0.023
## Adjusted R2                     0.013
## Residual Std. Error      177.722 (df = 100)
## F Statistic              2.364 (df = 1; 100)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Regression Analysis - Example 2

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                vcrime.pc
## ------------------------------------------------
## Year2021                        -54.111
##                                 (35.194)
##
## Constant                       424.337***
##                                 (24.886)
##
## ------------------------------------------------
## Observations                      102
## R2                               0.023
## Adjusted R2                      0.013
## Residual Std. Error        177.722 (df = 100)
## F Statistic                2.364 (df = 1; 100)
## ================================================
## Note:                  *p<0.1; **p<0.05; ***p<0.01
```

- $\widehat{\alpha} = 424.337 = E(vcrime.pc|2001) = E(vcrime.pc|Year2021 = 0)$
- $\widehat{\beta} = -54.111 = E(vcrime.pc|2021) - E(vcrime.pc|2001) = E(vcrime.pc|Year2021 = 1) - E(vcrime.pc|Year2021 = 0)$

Regression Analysis - Example 3

We now estimate a regression model which includes a constant and the number of law enforcement officers (per 100,000), `law.officer.pc`, as an explanatory variable (only 2021 data).

$$vcrime.pc_i = \alpha + \beta\ law.officer.pc_i + u_i$$

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                vcrime.pc
## ------------------------------------------------
## law.officers.pc                 0.204
##                                (0.193)
##
## Constant                      311.421***
##                                (61.296)
##
## ------------------------------------------------
## Observations                     50
## R2                               0.023
## Adjusted R2                      0.002
## Residual Std. Error       156.747 (df = 48)
## F Statistic              1.114 (df = 1; 48)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Regression Analysis - Example 3

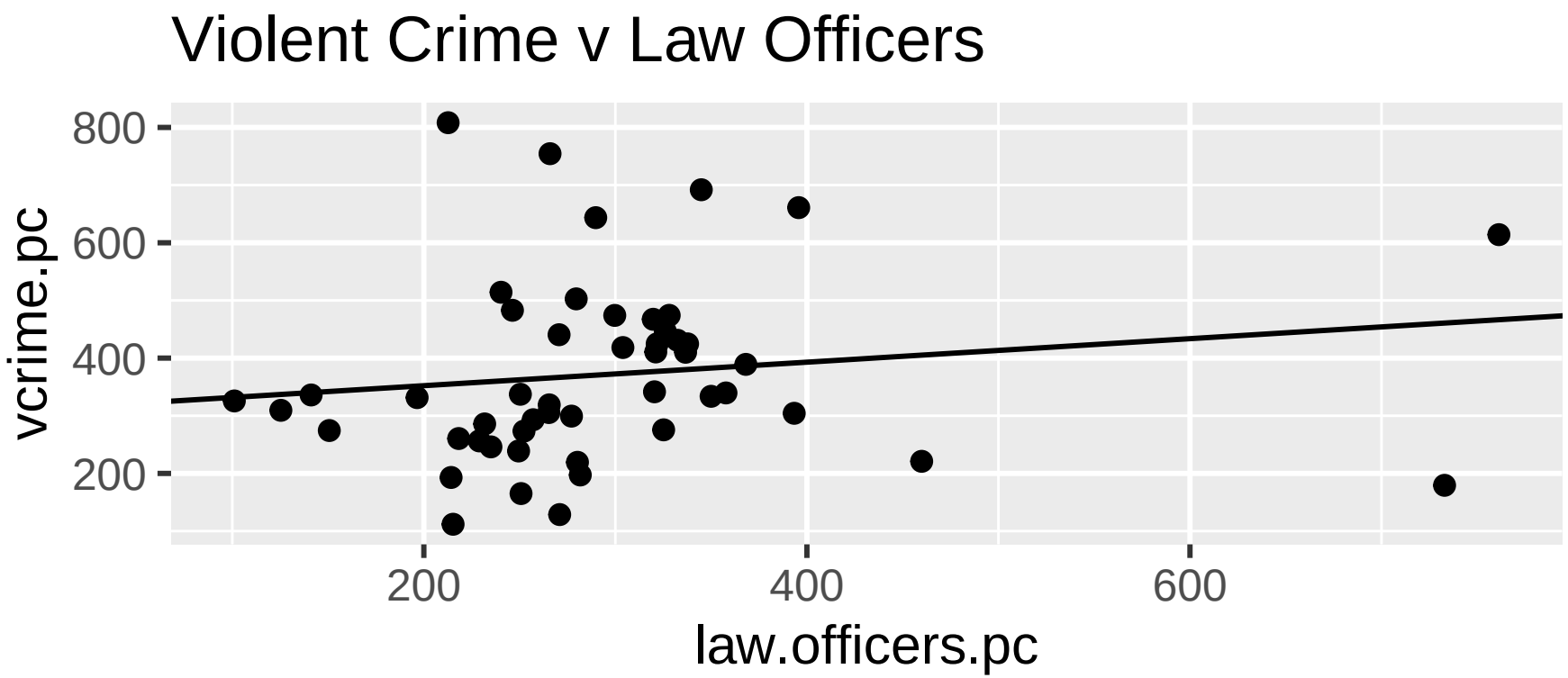
```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                vcrime.pc
## ------------------------------------------------
## law.officers.pc                 0.204
##                                (0.193)
##
## Constant                      311.421***
##                                (61.296)
##
## ------------------------------------------------
## Observations                     50
## R2                              0.023
## Adjusted R2                     0.002
## Residual Std. Error       156.747 (df = 48)
## F Statistic              1.114 (df = 1; 48)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

- $\widehat{\beta} = 0.235.$
- $\widehat{\alpha} = 311.421.$

Regression Analysis - Example 3

Let's present a graphical representation.

```
ggplot(test_data, aes(x=law.officers.pc, y=vcrime.pc)) +
    geom_point() +
    geom_abline(intercept = mod1$coefficients[1], slope = mod1$coefficients[2])+
    ggtitle("Violent Crime v Law Officers")
```



Regression Analysis - What does it actually do?

Two interpretations (note that here y is the dependent and x the explanatory variable)

1) Finds the regression line (via $\widehat{\alpha}$ and $\widehat{\beta}$) that the residual sum of squares $\Sigma(y_i - \widehat{\alpha} - \widehat{\beta}\, x_i)^2$. $\rightarrow$ Ordinary Least Squares (OLS)
2) Finds the regression line (via $\widehat{\alpha}$ and $\widehat{\beta}$) that ensures that the residuals ($\widehat{u}_i = y_i - \widehat{\alpha} - \widehat{\beta}\, x_i$) are with the explanatory variable(s).

In many ways 2) is the more insightful one.

Regression Analysis - What does it actually do?

$y = \alpha + \beta\ x + u$

Assumptions

One of the regression assumptions is that the (unobserved) error terms u are uncorrelated with the explanatory variable(s), x. Then we call x exogenous.

This implies that $Cov(x, u) = Corr(x, u) = 0$

In sample

$y_i = \quad \widehat{\alpha} + \widehat{\beta}\ x_i \quad + \widehat{u}$

Where $\widehat{\alpha} + \widehat{\beta}\ x_i$ is the regression-line.

In sample $Corr(x_i, \widehat{u}_i) = 0$ (is ALWAYS TRUE BY CONSTRUCTION).

Regression Analysis - Underneath the hood?

$y = \alpha + \beta\ x + u$

What happens if you call

`mod1 <- lm(vcrime.pc~law.officers,data=test_data)`?

You will recall the following from Year 1 stats:

$$\begin{aligned} \hat{\beta} &= \frac{\widehat{Cov}(y,x)}{\widehat{Var}(x)} = \frac{\widehat{Cov}(vcrime.pc, law.officers.pc)}{\widehat{Var}(law.officers.pc)} \\ \hat{\alpha} &= \overline{y} - \hat{\beta}\ \overline{x} = \overline{vcrime.pc} - \hat{\beta}\ \overline{law.officers.pc} \end{aligned}$$

The software will then replace $\widehat{Cov}(y,x)$ and $\widehat{Var}(x)$ with their sample estimates to obtain $\hat{\beta}$ and then use that and the two sample means to get $\hat{\alpha}$.

Regression Analysis - Underneath the hood?

Need to recognise that in a sample $\hat{\beta}$ and $\hat{\alpha}$ are really .

$$\begin{aligned}
\hat{\beta} &= \frac{\widehat{Cov}(y,x)}{\widehat{Var}(x)} \\
&= \frac{\widehat{Cov}(\alpha+\beta\ x+u,x)}{\widehat{Var}(x)} \\
&= \frac{\widehat{Cov}(\alpha,x)+\beta\widehat{Cov}(x,x)+\widehat{Cov}(u,x)}{\widehat{Var}(x)} \\
&= \beta\ \frac{\widehat{Var}(x)}{\widehat{Var}(x)}+\frac{\widehat{Cov}(u,x)}{\widehat{Var}(x)} = \beta+\frac{\widehat{Cov}(u,x)}{\widehat{Var}(x)}
\end{aligned}$$

So $\hat{\beta}$ is a function of the random term u and hence is itself a random variable. Once $\widehat{Cov}(y,x)$ and $\widehat{Var}(x)$ are replaced by sample estimates we get value which is draw from a

Regression Analysis - The Exogeneity Assumption

Why is **assuming** $Cov(x, u) = 0$ important when, in sample, we are guaranteed $Cov(x_i, \widehat{u}_i) = 0$?

If $Cov(x_i, u_i) = 0$ is **not true**, then

1) Estimating the model by OLS
2) The estimated coefficients $\widehat{\alpha}$ and $\widehat{\beta}$ are

3) The regression model has no

As we cannot observe u_i, the assumption of exogeneity cannot be tested and we need to make an argument using economic understanding.

Regression Analysis - Outlook

$y = \alpha + \beta\ x + u$

Much of empirical econometric analysis is about making the exogeneity assumption ($Corr(x, u) = 0$) more plausible/as plausible as possible. But this begins with thinking why an explanatory variable x is endogenous.

1) Most models have more than one explanatory variable.
2) Including more relevant explanatory variables can make the exogeneity assumption more plausible.
3) But fundamentally, if $Cov(u, x) = 0$ is implausible we need to find another variable z for which $Cov(u, z) = 0$ is plausible.

Outlook

Over the next weeks you will learn

- Simple OLS regression with dummy
- Endogeneity
- Multiple regression
- Difference-in-Difference (DiD) estimator