Introduction to the Design and Analysis of Experiments

Prof. Moacir Ponti
www.icmc.usp.br/~moacir

Instituto de Ciências Matemáticas e de Computação – USP

2018/1

Agenda

- Sampling
  - Sampling strategies
  - Explanatory and response variables
- Experiments
- Examining data and variables
- Hypothesis test

Sampling and Variables

Census vs Sampling

- It is rare to need a **census**
- **Sampling** is often sufficient if it is representative, but implies to accept some errors

Variables

- Numeric: discrete/continuous
- Categoric: ordinal/non-ordinal

Sampling and Variables

Exemple: check salt in a pan

- Exploratory analysis: **sampling** (why not a census?)

Sampling and Variables

Exemple: check salt in a pan

- Exploratory analysis: **sampling** (why not a census?)
- To conclude if needs more salt: **inference**

Sampling and Variables

Exemple: check salt in a pan

- Exploratory analysis: **sampling** (why not a census?)
- To conclude if needs more salt: **inference**
- We need a **representative** sampling, which requires **randomness**

Anecdotal evidence

- I met someone who was cured from asthma by homeopathy, so It must work.
- Testimonials on the Internet sais that garlic supplement helped some people to lose weight, so garlic should be an effective for weight-loss.
- My grandfather smoked and drank his whole life and lived until he was 95, so it is no unhealthy to drink and smoke.
- Today is 6°C, so global warming is a hoax.

Anecdotal evidence reliable? One man says "yes".

A STUDY CONDUCTED YESTERDAY by a man on himself concluded that self-reported anecdotal evidence is, in fact, both reliable and relevant.

The landmark study, conducted by Mark Mattingly of Virginia Beach in his apartment, concluded with 100% accuracy that data collected from personal experience can disprove other data conducted by reputable scientific institutions, thereby proving once and for all that "statistics can't be trusted".

In a press release Mr. Mattingly took aim at his detractors saying that "…this study shows what I've been telling people on the internet for years: all your fancy evidence and statistics don't mean nothing in the real world.".

Anecdotal evidence

It **is** based on data, however there are some issues

- Data only represent few cases
- It is not clear if those are representative
- Not necessarily the evidence is valid to falsify some claim!

Sampling bias

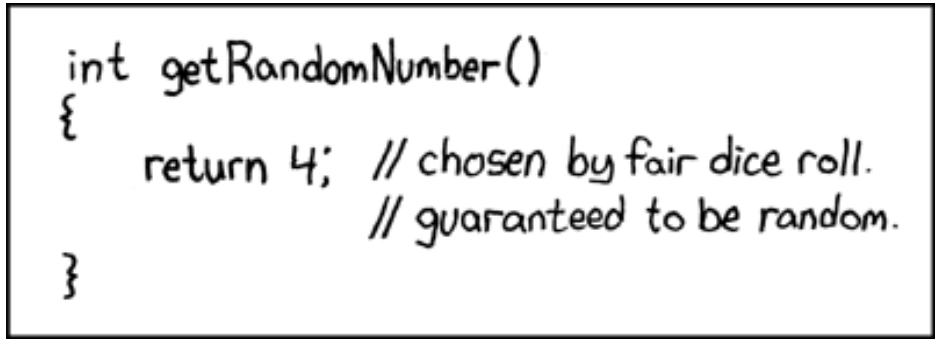
Convenience sample

Easily accesible sample

Non-response

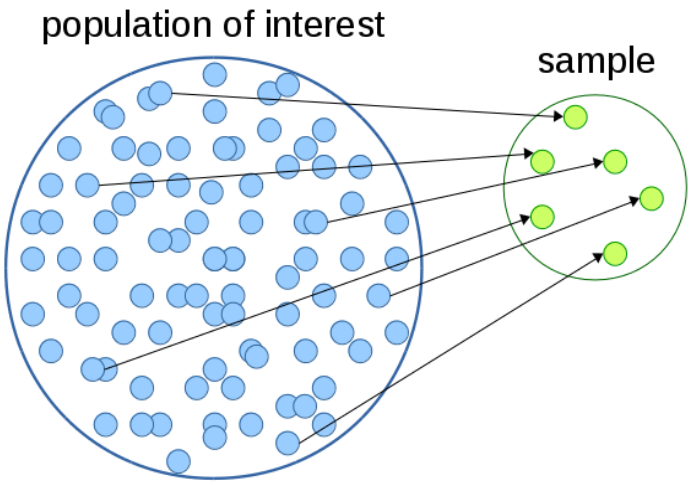
Only a fraction of a random sample responds or has interest on participating

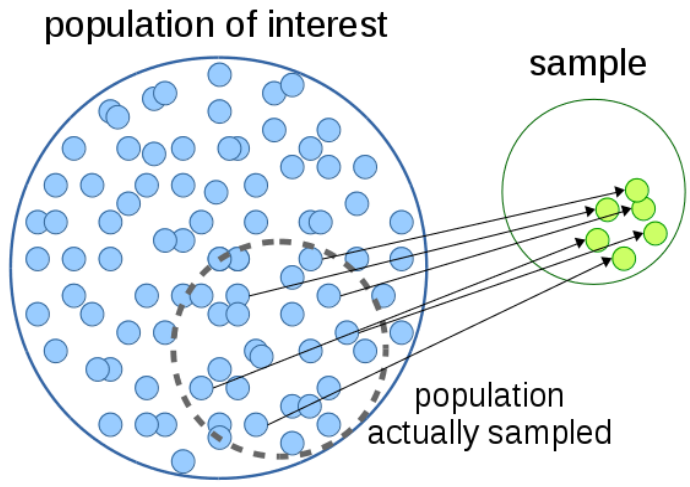
Sampling strategies



Thanks `http://xkcd.com`

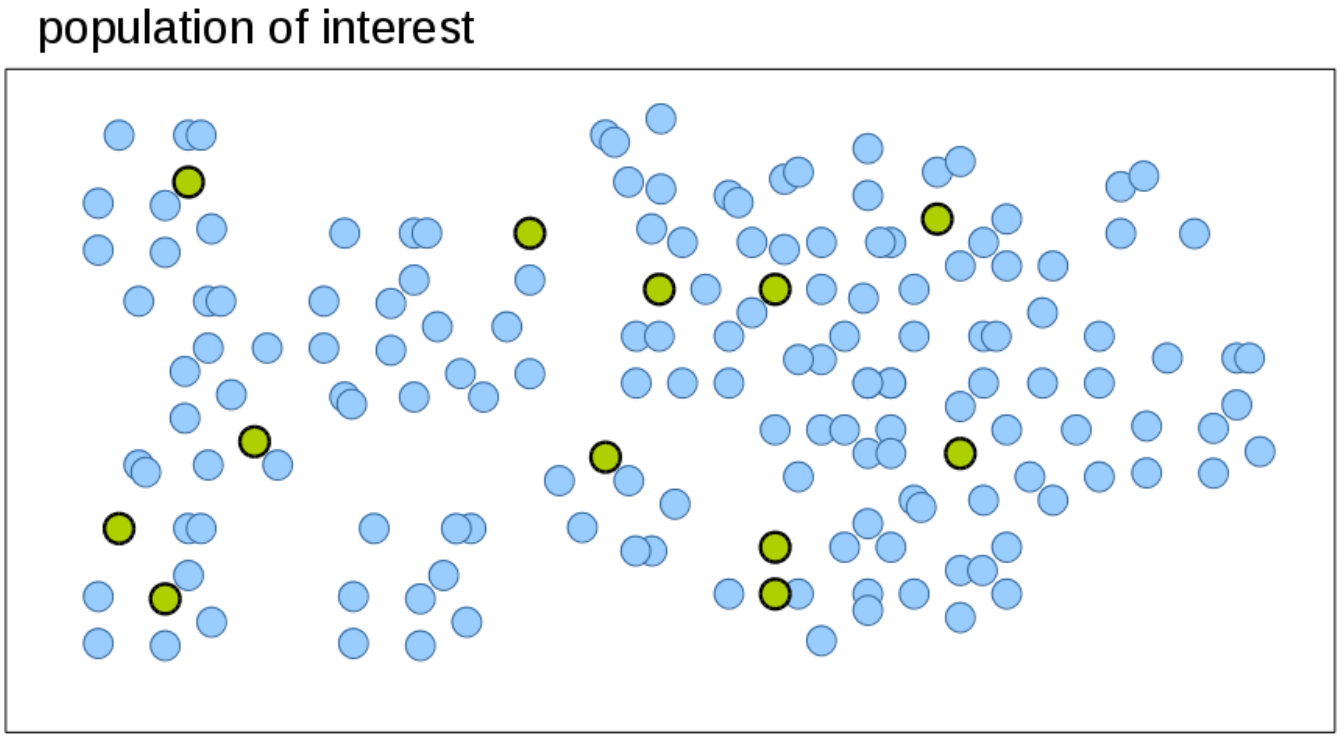
Sampling bias and i.i.d.





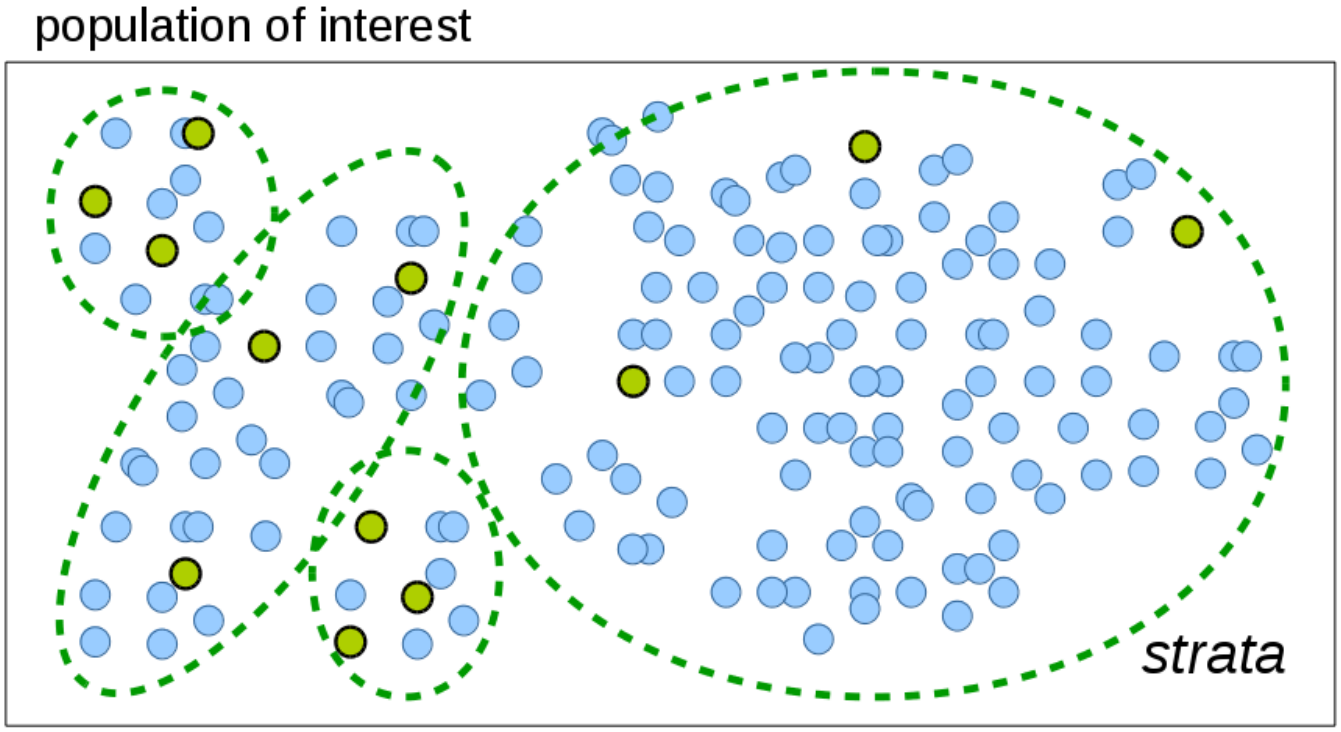
Sampling strategies

Simple random sampling



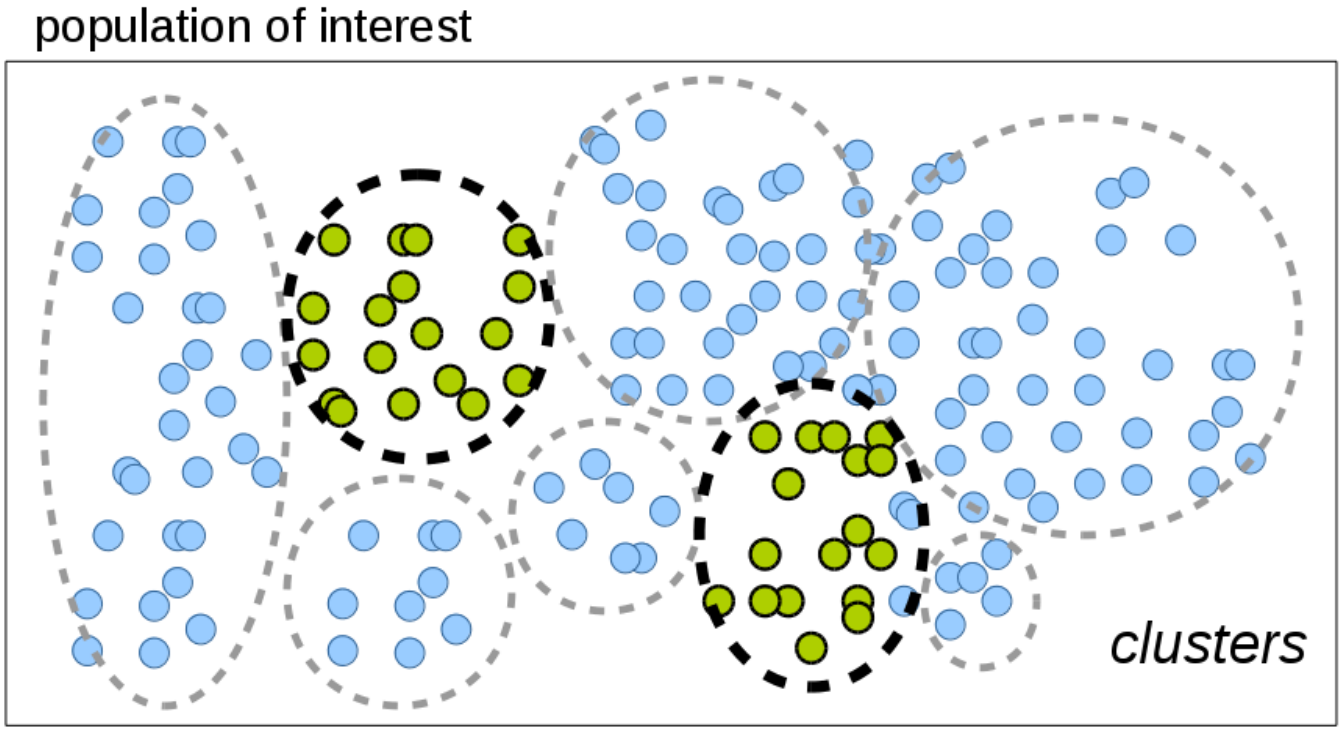
Sampling strategies

Estratified sampling



Sampling strategies

Clustering sampling



Explanatory and response variables

Question

- Is the classification accuracy of plants in images lower for natural images with higher levels of image compression?




Explanatory and response variables

Question

- Is the classification accuracy of plants in images lower for natural images with higher levels of image compression?




- Explanatory variable $\xrightarrow{\text{might affect}}$ Response Variable
- Independent variable $\xrightarrow{\text{might affect}}$ Dependent variable

Agenda

- Sampling
  - Sampling strategies
  - Explanatory and response variables
- Experiments
- Examining data and variables
- Hypothesis test

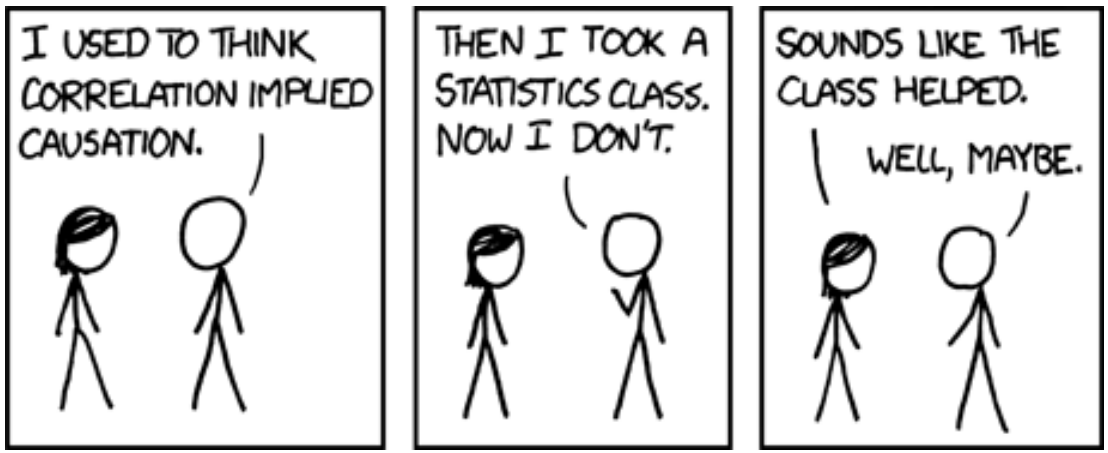
Experiments

Try to establish causal relations, correlations, or comparisons

1. **Control**: compare intervention with some control group,
2. **Randomization**: remove bias by experimenting over a randomized set of exemples (e.g. used for training/test, used to tune parameters and validate),
3. **Replication**: the more cases are observed, the more accurate are the estimates (e.g. cross-validation, repeated subsampling,etc.)
4. **Blocking**: evaluate some method in different blocks/scenarios, in a separate way.

Also common in experiments, but less common in computer science: placebo, placebo effect, blind / double-blind.

Causalidade vs Correlação



Thanks `http://xkcd.com`

Agenda

- Sampling
  - Sampling strategies
  - Explanatory and response variables
- Experiments
- Examining data and variables
- Hypothesis test

Measures and transformations

Measures of center and dispersion

- Mean and standard deviation
- Median and Interquartile Range (IQR)

Transformation

- Logarithm, Exponential, Squared-Root
- Normalization

Examples:

```
library(MASS)
data(cars)
data(cats)
data(pressure)
```

Measures and transformations

```
cars_o <- data.frame(speed=c(19,19,20,20,20), dist=c(190, 206, 210, 220, 238))
cars2 <- rbind(cars, cars_o)

# statistics
mean(cars2$dist)
sd(cars2$dist)

# robust statistic
median(cars2$dist)
IQR(cars2$dist)

plot(cars2)       # original data
plot(log(cars2))  # log
plot(sqrt(cars2)) # sqrt

plot(pressure)       # original data
plot(log(pressure))  # log transformation
```

Result analysis

Plots

- Scatterplots
- Boxplots

Example:

```
boxplot(cats$Bwt ~ cats$Sex)

cats_o <- data.frame(Sex=c('M','M','F'), Bwt=c(1.1,1.5,
cats2 <- rbind(cats, cats_o)

boxplot(cats2$Bwt ~ cats2$Sex)
```

Linear regression

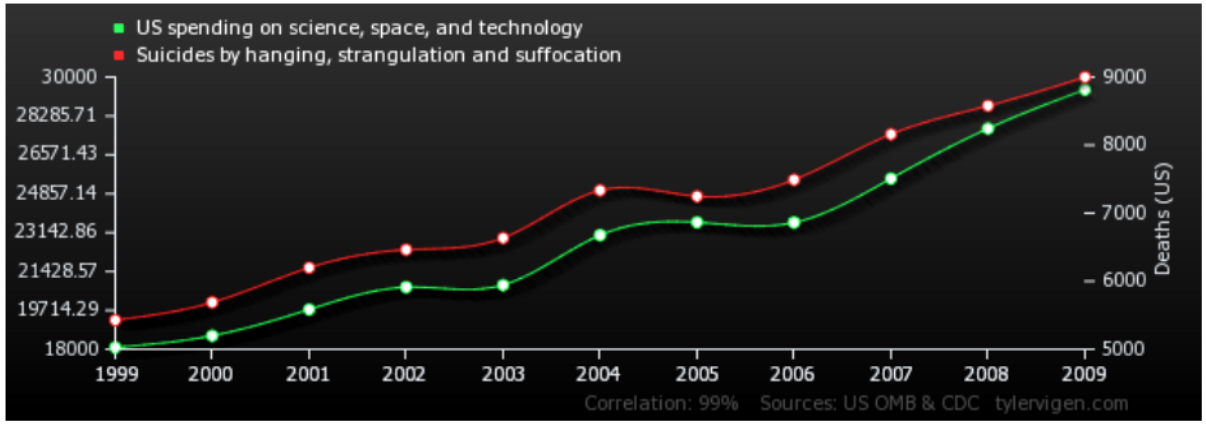
Fits a line on datapoints coming from two variables: one dependent, and one ore more independent.
If correlation is $\neq 0$, then: 1) A causes B, 2) B causes A, 3) some variable C causes A and B, 4) A causes C that causes B, or 5) correlation between A and B is coincidental.

```
model1 <- lm(cars$dist ~ cars$speed)
summary(model1)
plot(cars)
abline(model1)

model2 <- lm(cars2$dist ~ cars2$speed)
summary(model2)
plot(cars2)
abline(model2)
```

Correlation and Linear Regression

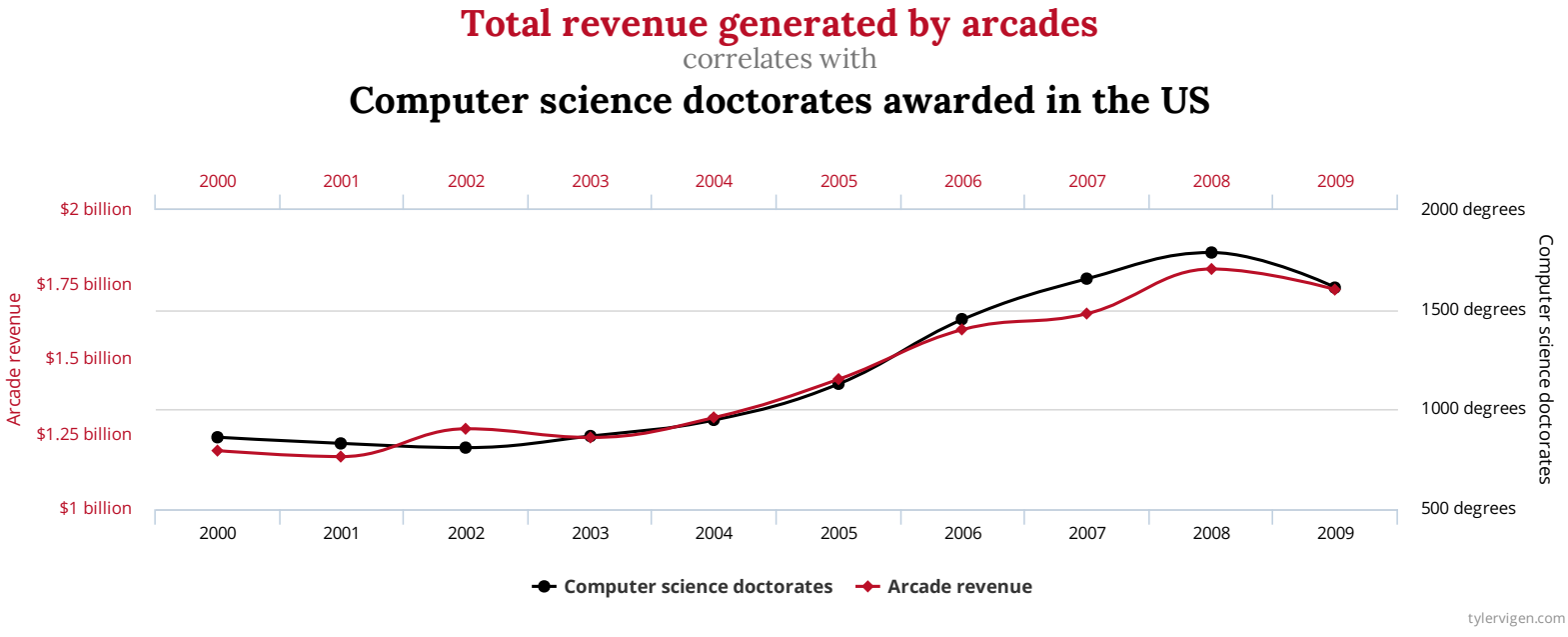
US spending on science, space, and technology
correlates with
Suicides by hanging, strangulation and suffocation



	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
US spending on science, space, and technology *Millions of todays dollars (US OMB)*	18,079	18,594	19,753	20,734	20,831	23,029	23,597	23,584	25,525	27,731	29,449
Suicides by hanging, strangulation and suffocation *Deaths (US) (CDC)*	5,427	5,688	6,198	6,462	6,635	7,336	7,248	7,491	8,161	8,578	9,000
Correlation: 0.992082											

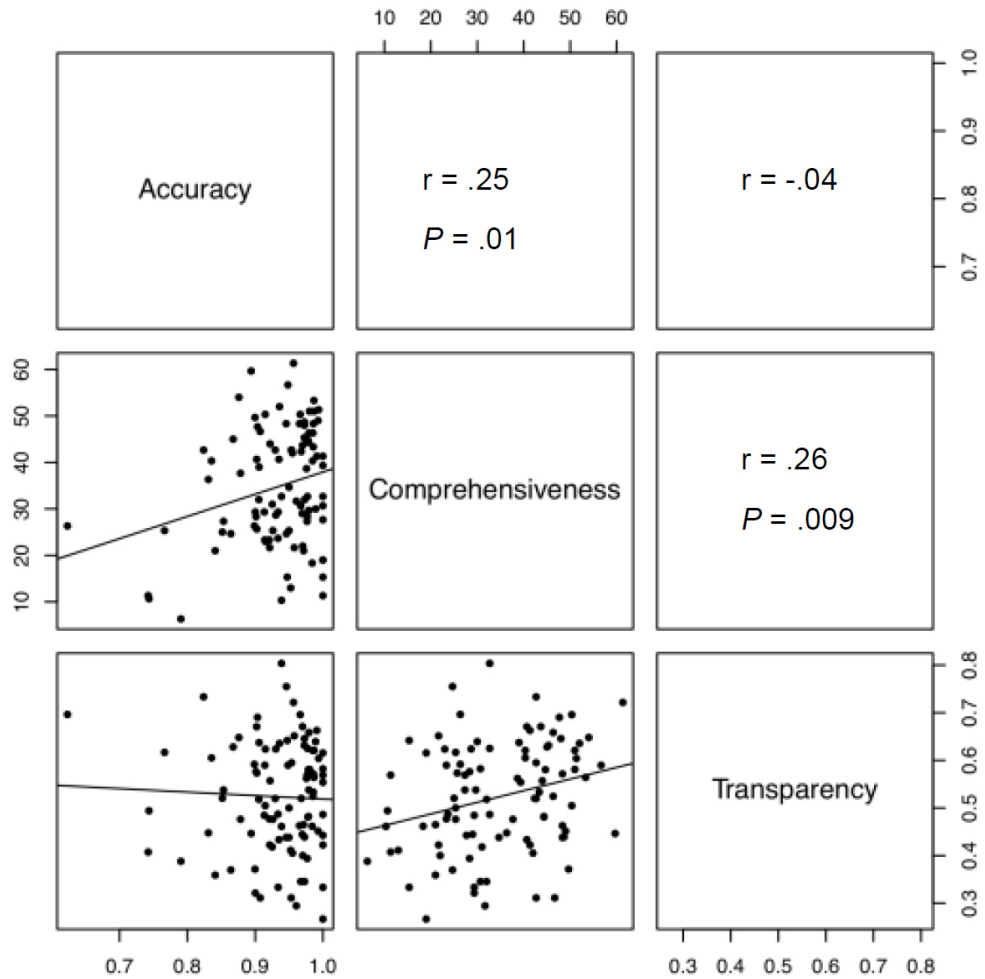
Thanks `http://tylervigen.com/`

Correlation and Linear Regression



Thanks `http://tylervigen.com/`

Correlation and Linear Regression



OBS: for $r = 0.25$, squared correlation is $R^2 = 0.06$

McNally et al (2012) http://www.i-jmr.org/2012/1/e1/

Agenda

- Sampling
  - Sampling strategies
  - Explanatory and response variables
- Experiments
- Examining data and variables
- Hypothesis test

Hypothesis test

1. Specifies **null hypothesis** and **alternative hypothesis**
2. Assumes null hypothesis is **true** and compute **test statistic**
3. Computes **p-value**: if null hypothesis is true, what is the probability of observing some as extreme as those we have?
   - if p is below some threshold α (which is the probability of error type I), the null hypothesis is rejected;
   - otherwise, do not reject null hypothesis.

Hypothesis test

- t-Student: for data under normal distribution;
- Wilcoxon: non-parametrical, uses rankings

- ANOVA: analyzes multiple sets via F statistics.
- Kruskal-Wallis: non-parametric version

Hypothesis test

Statisticians issue warning over misuse of P values

"Misuse of the P value — a common test for judging the strength of scientific evidence — is contributing to the number of research findings that cannot be reproduced..."

`http://www.nature.com/news/`

`statisticians-issue-warning-over-misuse-of-p-values-1.19503`