SOC542 Statistical Methods in Sociology II

Introduction and Review

Dr. Thomas Davidson

Rutgers University

January 27, 2025

Plan

- Introductions
- Course outline
- Review
  - Notation
  - Statistics
- Lab: R, RStudio, and statistics refresher

Learning goals

- The use of statistical analysis in empirical social science research
- Multiple regression
  - Conceptual and statistical background
  - Implementation in R
  - Interpretation
  - Violations assumptions and robustness
  - Frequentist and Bayesian approaches to estimation
- Proficiency in data handling, analysis, and visualization using R

Course outline

Structure

1. OLS regression (Weeks 1-4)
2. Non-linear variables and interactions (5-6)
3. Model checking and missing data (7)
4. Generalized linear models (8-11)
5. Clustered data (12)
6. Causal inference with observational data (13)
7. Presentations (14)

Course outline

Key themes

- Theory-driven statistical modeling
  - Causal, scientific models to guide implementation of statistical analyses

Course outline

Key themes

- Multiple approaches to estimation and inference
  - Quantification of uncertainty via Bayesian inference
  - Reduced emphasis on p-values

Course outline

Key themes

- Data-driven pedagogy
  - Preference for code and simulations over mathematical formalism

Course outline

Key themes

- Robust inferences
  - Appropriately specified statistical models
  - Model checking and comparison

Course outline

Key themes

- Reproducibility and transparency
  - Emphasis on reproducible and transparent analytic procedures

Course outline

Key themes

- Scientific communication
    - Clear, precise summaries of statistical models using tables and visualizations

Course outline

Key themes

- Theory-driven statistical modeling
- Multiple approaches to estimation and inference
- Data-driven pedagogy
- Robust inferences
- Reproducibility and transparency
- Scientific communication

Course outline

Assessment

- Homework assignments (40%)
  - Four assignments, each worth 10%
- Final paper (50%)
  - Phase 1: Research question, data, and methodology
  - Phase 2: Preliminary results
  - Phase 3: Submit final paper
- Presentations (10%)
  - Present results from final paper

Course outline

Required readings

- *Regression and Other Stories* by Andrew Gelman, Jennifer Hill, and Aki Vehtari
  - Main textbook, covers applied OLS regression and generalized models in R
- *Bayes Rules! An Introduction to Applied Bayesian Modeling* by Johnson, Alicia A., Miles Q. Ott, Mine Dogucu. 2021.
  - More detailed introduction Bayesian inference with examples in R

Course outline

Recommended readings

- *Causal Inference: The Mixtape* by Scott Cunningham
  - More formal econometric background and overview of causal approaches
- *Statistical Rethinking*, 2nd ed., by Richard McElreath
  - Supplementary textbook, provides additional material and deepens understanding of Bayesian inference and modern statistics
  - Recommended: McElreath's YouTube lecture series (2023 series currently underway)
- *R for Data Science* by Hadley Wickham and Garrett Grolemund.
  - A useful reference for data manipulation in R via the tidyverse
- *Data Visualization: A Practical Introduction* by Kieran Healy.
  - Great introduction to data visualization using R and ggplot

Notation review

Vectors

- A vector is a sequence of numbers
  - e.g. We take the heights of everyone in the class and record them in vector v

$$v = \{6.1, 5.9, 5.7, 6.0, 6.2, ..., 5.9\}$$

- We typically arrange these vertically as columns in a data table.
- We can use *indexing* to reference specific elements of the vector
  - v_i refers to arbitrary element i
  - v_1 indexes the first element, 6.1, and so on

Notation review

Summation

- The uppercase letter Σ is used to denote a summation. We can use it here to take the sum of all the values in vector v, where k is the length of the vector.

$$\sum v_i = \sum_{i=1}^{k} v_i = v_1 + v_2 + ... + v_k$$

- We can compute sums in R using the `sum()` function, where the thing we are summing over is included in the parentheses, e.g. `sum(v)`.

Notation review

Products

- The uppercase letter Π is used to denote the product operation. *We will encounter it far less frequently than summation.*

$$\prod v_i = \prod_{i=1}^{k} v_i = v_1 * v_2 * \ldots * v_k$$

Notation review

Matrices

- We often want to represent multiple vectors as a matrix.
- Let's say we also collected each students' age, we could represent the ages as a vector u.
  - v and u can be combined together in a matrix M (typically we use lowercase for vectors and constants and uppercase for matrices):

$$M = \left\{ \begin{matrix} 6.1 & 24 \\ 5.9 & 22 \\ 5.7 & 27 \\ 6.0 & 30 \\ 6.2 & 25 \\ ... & ... \\ 5.9 & 26 \end{matrix} \right\}$$

Notation review

Matrices

- We can index elements of a matrix in the following way
  - $M_{i,j}$ refers to the i^{th} row of column j
  - e.g. $M_{4,2}$ indexes the age of the 4^{th} student

Notation review

Vectors and matrices in R

```
v <- c(1,2,3)
print(sum(v))
```

```
## [1] 6
```

```
print(prod(v))
```

```
## [1] 6
```

Notation review

Vectors and matrices in R

We can use `cbind` to combine vectors columnwise into a matrix.

```
u <- c(1,1,1)
M <- cbind(v,u)
print(M[3,1]) # M[row, column]
```

```
## v
## 3
```

```
print(M)
```

```
##      v u
## [1,] 1 1
## [2,] 2 1
## [3,] 3 1
```

Notation review

Vectors and matrices in R

We can *transpose* this matrix using `t` if we want to treat these columns as rows.

```r
print(t(M))
```

```
##   [,1] [,2] [,3]
## v    1    2    3
## u    1    1    1
```

Statistics review

Random variables

- The value of a random variable depends on the outcome of random events
  - e.g. x is a random variable representing the outcome of a series of coin tosses

Statistics review

Probability distributions

- Random variables are drawn from probability distributions
  - The probability of tossing a coin and getting a head is a defined by the Bernoulli distribution
  - The number of heads in a sequence of coin tosses is defined by the binomial distribution
  - The height of a randomly chosen adult male is drawn from a normal distribution

Statistics review

Probability distributions

- Distributions are defined by parameters that modify their shape and scale.
  - A Bernoulli distribution has a single parameter, p
  - A binomial distribution has two parameters, n and p
  - A normal distribution has two parameters, a mean μ and a standard deviation of σ
    - Often we refer to this using the shorthand $N(\mu, \sigma)$

Probability distributions in R

A random variable drawn from a binomial distribution can take a value of 0 or 1, where p is the probability of a 1. $p = 0.5$ is equivalent to a fair coin toss. The Bernoulli distribution is a special case where $n = 1$.

```
x <- rbinom(n=1, 1, 0.5)
print(x)
```

```
## [1] 0
```

Statistics review

Probability distributions in R

In this case, I make 10 draws from a binomial distribution, simulating ten tosses of a fair coin.

```
x <- rbinom(10, 1, 0.5)
print(x)
```

```
##  [1] 0 1 0 0 1 1 0 0 1 0
```

Statistics review

Probability distributions in R

A Normal, or Gaussian, distribution is a continuous distribution with two parameters. The standard normal distribution has a mean $\mu = 0$ and standard deviation $\sigma = 1$. Here is a single random draw:

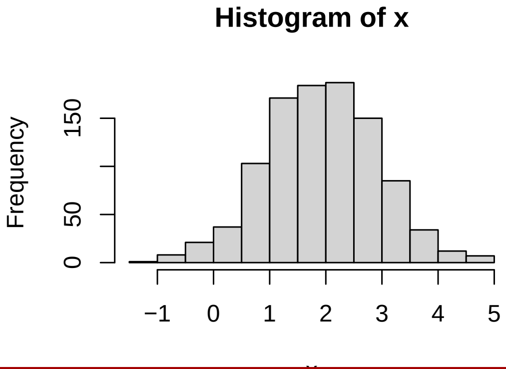
```r
x <- rnorm(1, mean = 0, sd = 1)
print(x)
```

```
## [1] 1.477438
```

Statistics review

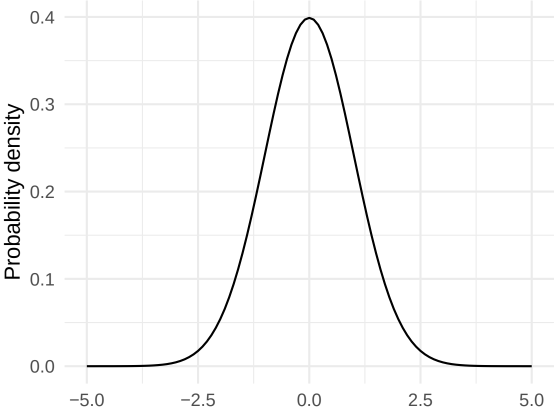
Probability distributions in R

In this case, we can make 1000 draws from a normal distribution with a mean of 2 and a standard deviation of 1. Since there are a lot of values it is best to plot them using a histogram.



Probability distributions in R

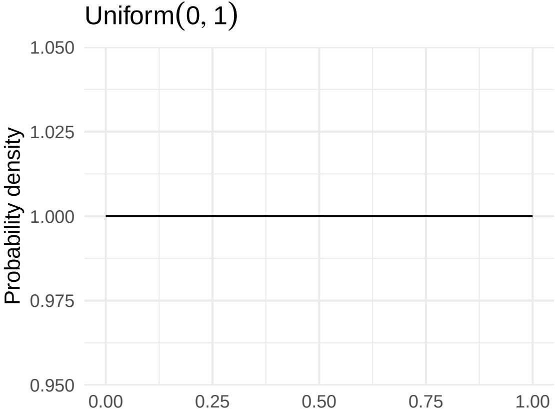
A *probability density function* shows the probability of observing particular random draws from a given distribution.



Statistics review

Probability distributions in R

In a uniform distribution, every value in the interval $[a, b]$ is equally likely.



Statistics review

Expected values

- The expectation of a random variable is denoted by $\mathbb{E}[x]$. It is the long-run average of the random variable over many repeated trials.
- The expected value of a constant $x = c$ is c. E.g. $\mathbb{E}[2] = 2$
- The expected value of a random variable is its mean.
  - e.g. If x represents a vector of values drawn from a normal distribution, our best guess as to the value of any one realization x_i is μ.
- We can use expectations notation to represent relationships between variables, e.g. $\mathbb{E}[y|x]$

Statistics review

Expectations as weighted averages

- Discrete case, where p_i is the probability of observing a particular value of x

$$\mathbb{E}[x] = \Sigma_{i=1}^{k} x_i p_i$$

- In the continuous case, where $f(x)$ is a probability density function, the expectation is an integral over the possible values of x.

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} x f(x) dx$$

Statistics review

Populations and samples

- Classical statistics is based upon the assumption that our observations x are drawn from an underlying population.
- In a simple random sample, we draw n instances of a random variable from the population
  - These draws are assumed to be *independent* and *identically distributed* (IID)
- For example, a hypothetical population might be adults residing in the United States and a sample would be a randomly sampled subset of these adults.

Statistics review

Means

- The population mean value of a random variable is defined in the following manner

$$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{\sum_{i=1}^{N} x_i}{N}$$

- The sample mean $\bar{x}$ is defined by the following equation. n is lowercase to denote $n \subseteq N$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{\sum_{i=1}^{n} x_i}{n}$$

Statistics review

Variance

- Variance is the average of the squared deviations from the mean. For the population it is defined as

$$\sigma_x^2 = \frac{1}{N}\Sigma(x_i - \mu)^2$$

- The sample variance includes a degrees of freedom correction.

$$\sigma_{\bar{x}}^2 = \frac{1}{n-1}\Sigma(x_i - \bar{x})^2$$

Statistics review

Standard deviation

- Variance is difficult to interpret. The standard deviation is a scaled measure of variance, typically denoted using σ. In the population, it is equal to

$$\sigma_x = \sqrt{\sigma_x^2} = \sqrt{\frac{1}{N}\Sigma(x_i - \mu)^2}$$

- The sample standard deviation is thus

$$\sigma_{\bar{x}} = \sqrt{\sigma_{\bar{x}}^2} = \sqrt{\frac{1}{n-1}\Sigma(x_i - \bar{x})^2}$$

Statistics review

The sampling distribution of the mean

- The mean of an IID random sample is distributed according to a *sampling distribution*
- It has the following properties:

$$\mathbb{E}[\bar{x}] = \mu_x$$

$$var(\bar{x}) = \sigma^2_{\bar{x}} = \frac{\sigma^2_x}{N}$$

Statistics review

Simulating the sampling distribution

We can simulate the sample mean of 1000 normal distributions with the same population parameters and compare the sample means to the expectations.

```
mu <- 100; sigma <- 10; n <- 100 # population parameters
sims <- replicate(1000, mean(rnorm(n, mu, sigma)))
print(round(mean(sims),2)) # E[mu] = 100
```

```
## [1] 99.99
```

```
print(round(var(sims),2)) # E[sigma^2/n] = (10^2)/100 = 1
```

```
## [1] 0.9
```

Statistics review

The Law of Large Numbers

- When sample size is large, $\bar{x}$ is close to μ_x with a high probability
- A large IID sample can therefore approximate the sampling distribution
  - Such samples are *asymptotic* because approximations become exact in the limit as $n \to \infty$
- Under such conditions, $\bar{x}$ is considered a *consistent* estimator for μ_x

Statistics review

The Law of Large Numbers

The sample mean from a large IID sample closely approximates the population mean.

```
large.sample <- rnorm(1e6, mu, sigma)
print(mean(large.sample))
```

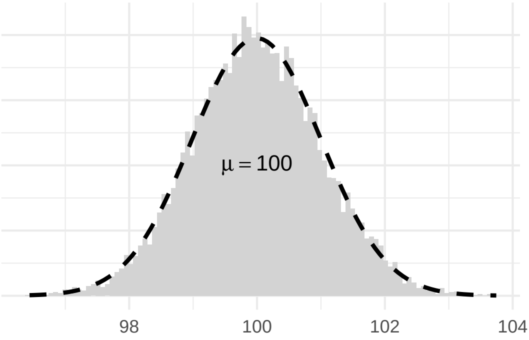
```
## [1] 99.99705
```

The Central Limit Theorem

- CLT: The distribution of $\bar{x}$ is well approximated by a normal distribution when n is large
  - This is approximately true even if x are not normally distributed

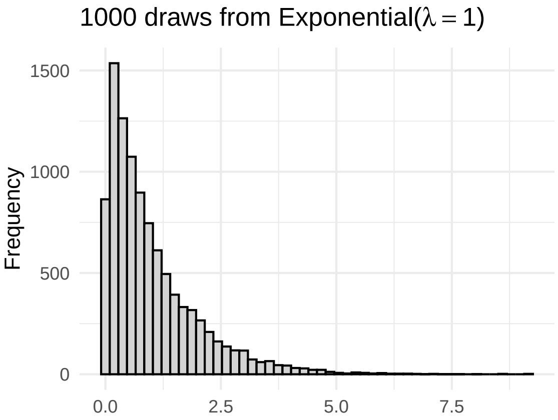
Statistics review

$N(\mu, \sigma^2)$

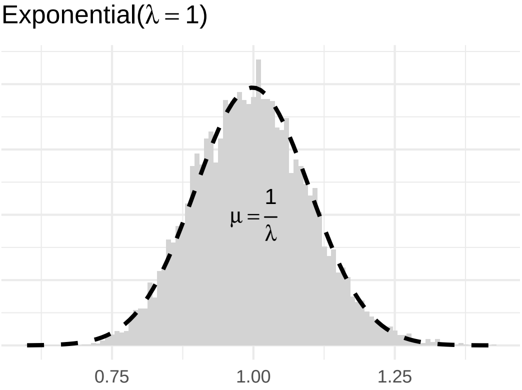


Distribution of 10,000 sample means, where n = 100

Statistics review



Statistics review



Distribution of 10,000 sample means, where n = 100.

Statistics review

Standard error of the sample mean

- The standard error of the sample mean is defined as the sample standard deviation divided by the square root of n.

$$SE_{\bar{x}} = \frac{\sigma_{\bar{x}}}{\sqrt{n}}$$

- The standard error is used to communicate *uncertainty* since we cannot observe the true population mean μ but only the sample mean $\bar{x}$.
- Theoretically, the standard error is the *standard deviation of the sampling distribution*.

Statistics review

Standard error as standard deviation of the sampling distribution

We can show that a standard error of a large random sample is a good approximate of the standard deviation of the sampling distribution.

```
N <- 10000
s.dist <- replicate(1000, mean(rnorm(N)))
print(round(sd(s.dist),3))
```

```
## [1] 0.01
```

```
x <- rnorm(N)
print(round(sd(x)/sqrt(N),3))
```

```
## [1] 0.01
```

Statistics review

Estimating the standard error of the mean in R

```
mu <- 10 # population mean
sigma2 <- 1 # population standard deviation
N <- 100
x <- rnorm(N, mu, sigma2)
print(mean(x)) # sample mean
```

```
## [1] 9.862993
```

```
print(sd(x)/sqrt(N)) # sample SE
```

```
## [1] 0.09972622
```

Statistics review

Confidence intervals

- The standard error is used to define a confidence interval. By convention, a 95% confidence interval has lower and upper bounds of

$$[\bar{x} - 1.96SE_{\bar{x}}, \bar{x} + 1.96SE_{\bar{x}}]$$

- When n is large, the 95% confidence interval around $\bar{x}$ contains the true value μ_x in 95% of all possible random samples.

Statistics review

Confidence intervals in R

We can use the results of the previous simulation to compute confidence intervals.

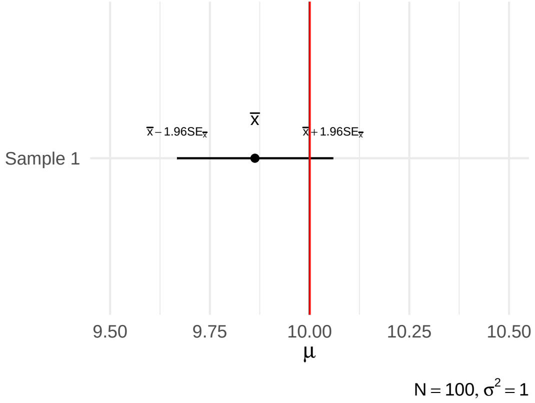
```
xbar <- mean(x)
SE <- sd(x)/sqrt(N)
lower <- xbar-1.96*SE
print(lower)
```

```
## [1] 9.667529
```

```
upper <- xbar+1.96*SE
print(upper)
```

```
## [1] 10.05846
```

Confidence intervals in R



Statistics review

One-sample t-tests

- We can use a one-sample t-test to assess whether there is a statistically significant difference between our sample mean and a null hypothesis.
- The *test statistic t* is obtained using the following equation:

$$t = \frac{\bar{x} - \mu}{\frac{\sigma_{\bar{x}}}{\sqrt{n}}}$$

Statistics review

p-values and statistical significance

- In classical statistics, we use a test statistic to calculate a *p-value*.
  - This represents the probability of drawing a test statistic at least as large as the observed test statistic, assuming the null hypothesis is correct
- Smaller p-values indicate that a result is less likely to be due to chance
  - By convention, $p < 0.05$ is considered the threshold for statistical significance.

Statistics review

One-sample t-tests in R

```r
t.test(x, mu=10)
```

```
##
##  One Sample t-test
##
## data:  x
## t = -1.3738, df = 99, p-value = 0.1726
## alternative hypothesis: true mean is not equal to 10
## 95 percent confidence interval:
##   9.665114 10.060871
## sample estimates:
## mean of x
##  9.862993
```

Statistics review

One-sample t-tests in R

We can calculate the t-statistic and directly look up the corresponding p-value using a Student t distribution with N-1 degrees of freedom.[1]

```
t <- (xbar-mu)/(sd(x)/sqrt(N))
print(t)
```

```
## [1] -1.373835
```

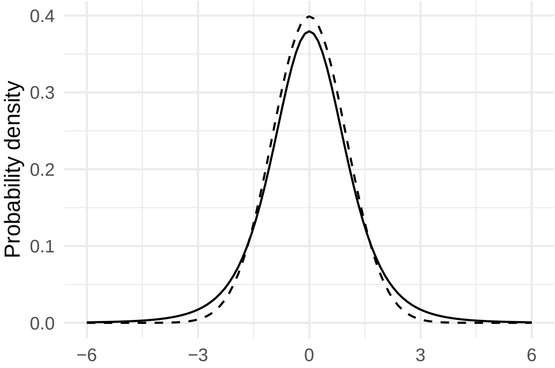
```
p <- pt(abs(round(t,5)), df = N-1, lower.tail = FALSE)
print(round(p*2,4)) # p*2 = two-tailed p-value
```

```
## [1] 0.1726
```

[1] Rounding ensures that the results are equivalent to the 't.test' function used above.

Statistics review

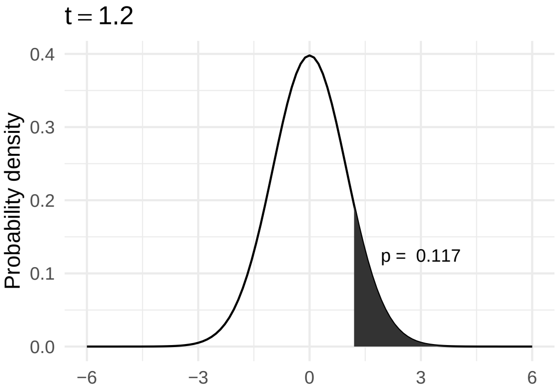
The t-test and the Student t distribution



Student t distribution with 5 d.f.
N(0,1) dashed line

Statistics review

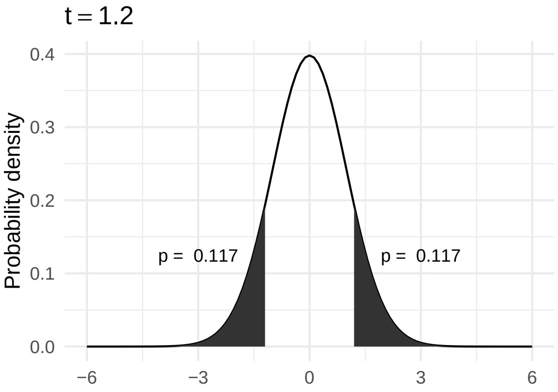
Area under one tail



Student t distribution with 99 degrees of freedom.

Statistics review

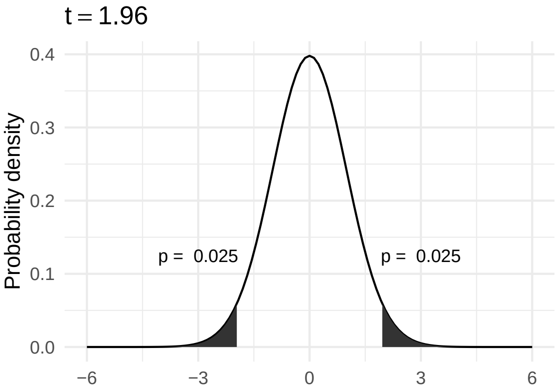
Area under both tails



Student t distribution with 99 degrees of freedom.

Statistics review

Area under both tails



Student t distribution with 999 degrees of freedom.

Statistics review

Random sampling and p-values

Another way to show the same result is to draw random variables from a Student t distribution and calculate the proportion that fall above the chosen significance threshold.

```
random.t <- rt(n = 10000, df = 1000-1)
round(length(random.t[abs(random.t) >= 1.96])/
        length(random.t),2)
```

```
## [1] 0.05
```

We can expect to see a t-statistic where $|t| \geq 1.96$ approximately 5% of the time purely due to chance.

Statistics review

Testing for differences in means

- Often we want to know whether the mean values of two random variables, μ_x and μ_y, are different from one another.
- We can test for this by computing calculating the difference between the sample means and the uncertainty about that difference
- The equation for the two-sample t-test is:

$$t = \frac{\mu_x - \mu_y}{SE(\mu_x - \mu_y)}$$

where

$$SE(\mu_x - \mu_y) = \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$$

Statistics review

Testing for differences in means

```r
n <- 1000
x <- rnorm(n, mean = 12)
y <- rnorm(n, mean = 11)
xbar <- mean(x)
ybar <- mean(y)
varx <- (1/(n-1))*sum((x-xbar)^2)
vary <- (1/(n-1))*sum((y-ybar)^2)
SE <- sqrt((varx/n)+(vary/n))
t <- (xbar-ybar)/SE
print(round(t,3))
```

```
## [1] 20.831
```

Statistics review

Testing for differences in means

```r
t.test(x, y)
```

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = 20.831, df = 1994.7, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not
## 95 percent confidence interval:
##  0.8753134 1.0572567
## sample estimates:
## mean of x mean of y
##  11.97403  11.00774
```

Statistics review

Type I and Type II errors

- A Type I (false positive) error occurs when we incorrectly reject the null hypothesis
  - e.g. In the one-sample test, we reject the hypothesis that $\bar{x} = 10$ given $\mu = 10$
- A Type II (false negative) error occurs when we incorrectly fail to reject the null hypothesis
  - e.g. If we did not have sufficient statistical power for the test above, we might fail to reject the null that $\bar{x} = \bar{y}$
- In classical statistics, if our chosen significance level is `p < 0.05` then we expect to see such errors approximately 5% of the time

Sign (S) and magnitude (M) errors

- Gelman, Hill and Vehtari draw attention to two kinds of errors that are often overlooked but can lead to profoundly incorrect inferences:
  - Sign errors: the sign of a relationship is incorrect
    - e.g. We observe a positive relationship between x and y when the true relationship is negative
  - Magnitude errors: the observed effect is severely over- or under-estimated
    - e.g. We think $\bar{x} = 10$ but $\bar{x} = 0.1$

Statistics review

Expected mean and variance of two random variables

- The expected mean of the sum of two random variables is

$$E[x+y] = E[x] + E[y] = \mu_x + \mu_y$$

- The expected variance is the sum of the variances plus twice their covariance

$$var(x+y) = var(x) + var(y) + 2cov(x,y)$$

- If x and y are independent then $cov(x,y) = 0$ and $var(x+y) = var(x) + var(y)$

Statistics review

Covariance

- Covariance is the a measure of the joint variability of two random variables
- The expectation of the covariance between x and y is

$$cov(x, y) = E[xy] - E[x]E[y]$$

- For a population, the covariance is

$$cov(x, y) = \frac{1}{N}\Sigma(x_i - \mu_x)(y_i - \mu_y)$$

- Sample covariance is defined as

$$cov(x, y) = \frac{1}{n-1}\Sigma(x_i - \bar{x})(y_i - \bar{y})$$

Statistics review

Correlation

- Correlation is a scaled version of covariance. We divide the covariance by the product of the standard deviations.

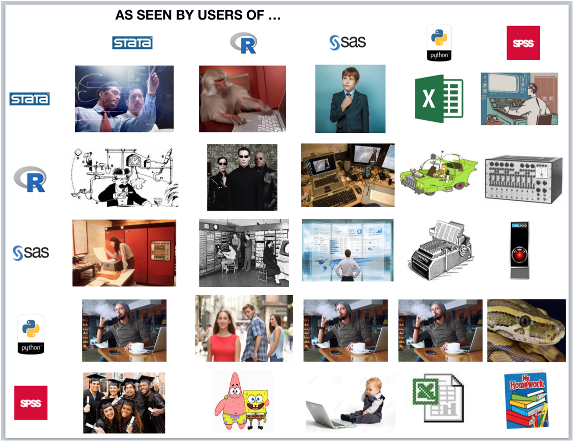
$$\rho(x,y) = \frac{\frac{1}{n-1}\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y} = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

- The Greek letter ρ is typically used to refer to correlation. The correlation coefficient ranges from -1 to 1.

Why R?

- Free and open-source
- Versatile
  - Statistical computing (`lm`, `glm`, ‘rstanarm“)
  - Data manipulation (`tidyverse`)
  - Visualization (`ggplot`)
  - General purpose programming
- Active developer community
  - Lots of packages with regular updates
  - New statistical and econometric packages

Why R?



Source: Kieran Healey

RStudio

Overview

- RStudio is an Integrated Development Environment for programming in R
  - Run code in the console or in scripts
  - Easy to view data, objects in memory, plots
  - Easy to create output such as papers or slides
  - Integration with Github

RMarkdown

Overview

- RMarkdown is an interactive coding environment
  - RMarkdown documents can combine text, LaTeX code, R code, and any output.
  - These slides are rendered using RMarkdown
  - You will be using RMarkdown for your homework assignments and hopefully your papers

Next lecture

- Ordinary Least Squares regression with two variables

Lab 1

- RStudio and RMarkdown
- Manipulating random variables
- Intro to `tidyverse` and `ggplot` for data manipulation and visualization