

SOC542 Statistical Methods in Sociology II

Introduction and Review

Dr. Thomas Davidson

Rutgers University

January 24, 2022

Plan

- ▶ Introductions
- ▶ Course outline
- ▶ Review
 - ▶ Notation
 - ▶ Statistics
- ▶ Lab: R and RStudio

Learning goals

- ▶ Proficiency in data handling, analysis, and visualization using R
- ▶ Multiple regression (OLS and GLMs)
 - ▶ Conceptual understanding
 - ▶ Implementation in R
 - ▶ Interpretation
 - ▶ Violations assumptions and robustness
- ▶ Understand and apply frequentist and Bayesian estimation methods
- ▶ Replicate, reassess, and extend cutting-edge quantitative research

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

Assessment

- ▶ Homework assignments (50%)
 - ▶ Simple OLS regression
 - ▶ Multiple regression
 - ▶ Model checking and missing data
 - ▶ Logistic regression
 - ▶ Predicting counts and categories

Course outline

Assessment

- ▶ Replication paper (40%)
 - ▶ Phase 1: Identify suitable paper for replication
 - ▶ Phase 2: Replicate a key result
 - ▶ Phase 3: Reassess the result
 - ▶ Phase 4: Extend the analysis
- ▶ Presentations (10%)

Course outline

Readings

- ▶ *Regression and Other Stories* by Andrew Gelman, Jennifer Hill, and Aki Vehtari
 - ▶ Main textbook, covers applied OLS regression and generalized models in R
- ▶ *Statistical Rethinking*, 2nd ed., by Richard McElreath
 - ▶ Supplementary textbook, provides additional material and deepens understanding of Bayesian inference
 - ▶ Recommended: McElreath's YouTube lecture series
- ▶ *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

Course outline

Policies

- ▶ Diversity and inclusion
- ▶ Code of conduct and academic integrity
- ▶ Accommodations
- ▶ COVID-19

See the syllabus for further details on each policy.

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 access specific elements of the vector
 - ▶ e.g. v_2 indexes the second element, 5.9

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.

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

- ▶ 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.

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

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):

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

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 4th student

Notation review

Vectors and matrices in R

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

```
u <- c(1,1,1)
```

```
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.

```
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.

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

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


Statistics review

Random variables

- ▶ A random variable can take on any value
 - ▶ x is a random variable
- ▶ A random variable could refer to a single value or a vector, where x_i indexes the i^{th} element

Statistics review

Probability distributions

- ▶ Random variables are drawn from probability distributions
 - ▶ The probability of tossing a coin and getting a head is 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 $N(\mu, \sigma)$

Statistics review

Probability distributions in R

We can easily use R to construct random variables drawn from different distributions. First, 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 and calculate the number of 1s.

```
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 distribution, often referred to as a Gaussian, is a continuous distribution with two parameters. The standard normal distribution has a mean $\mu = 0$ and standard deviation $\sigma = 1$

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

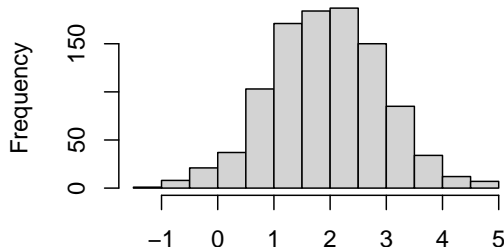
```
## [1] 1.477438
```

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.

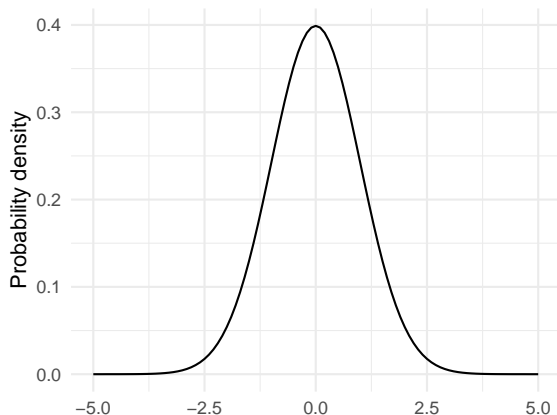
Histogram of x



Statistics review

Probability distributions in R

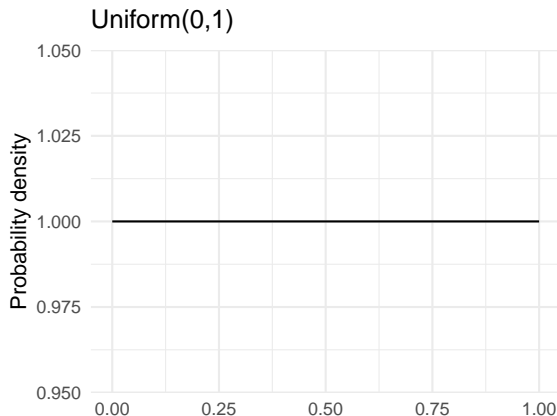
We can also plot the density of a normal distribution directly without drawing any random variables.



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 will see expectations more when we cover regression, 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

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

- ▶ 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 selected 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. Note that 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

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

- ▶ The sample variance includes a degrees of freedom correction.

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

Statistics review

Standard deviation

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

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

- ▶ The sample standard deviation is thus

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

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:

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

$$\text{var}(\bar{x}) = \sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{N}$$

Statistics review

Simulating the sampling distribution

We can draw five random samples from a normal distribution and compare the sample means to the expectations.

```
mu <- 100; sigma <- 10; n <- 100
sims <- replicate(5, mean(rnorm(n, mu, sigma)))
print(round(mean(sims),2)) #  $E[\mu] = 100$ 

## [1] 100.07

print(round(var(sims),2)) #  $E[\sigma^2/n] = 1$ 

## [1] 1.21
```

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 be used to approximate the sampling distribution
 - ▶ Such samples are *asymptotic* because approximations become exact in the limit as $n \rightarrow \infty$
- ▶ Under such conditions, \bar{x} is considered to be *consistent* 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] 99.99791
```

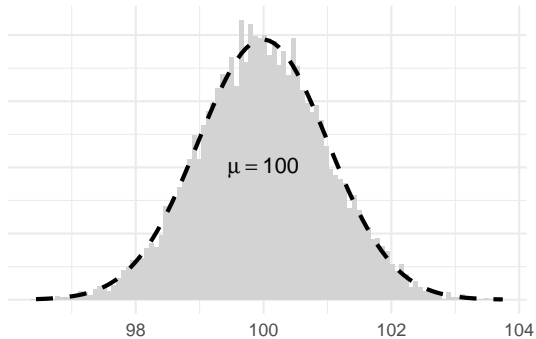
Statistics review

The Central Limit Theorem

- ▶ 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

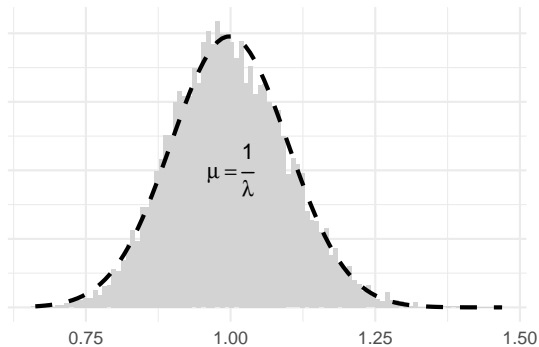
$N(\mu, \sigma^2)$



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

Statistics review

Exponential($\lambda = 1$)



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, it 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

We can draw from a normal distribution in R and estimate the

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

## [1] 9.885782

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

## [1] 0.08159911
```

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, a 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 easily use the previous simulation to compute confidence intervals.

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

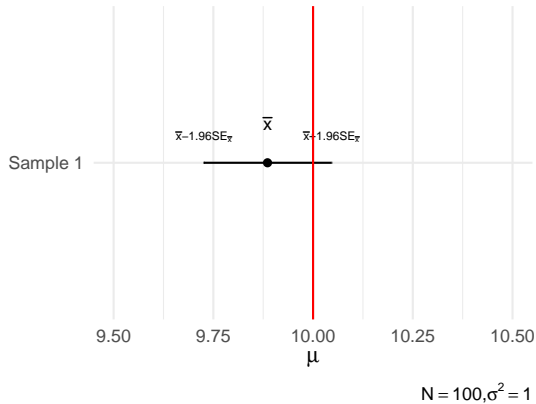
```
## [1] 9.725848
```

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

```
## [1] 10.04572
```

Statistics review

Confidence intervals in R



Statistics review

One-sample t-tests

- ▶ In classical statistics we are often interested in testing hypotheses. 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.
- ▶ In the example above we know that the population mean $\mu = 10$. We can test whether the sample mean is statistically significantly different from the population mean using a *t-test*:

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

Statistics review

p-values and statistical significance

- ▶ The t statistic is an example of a *test statistic*
- ▶ 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 to be the threshold for statistical significance

Statistics review

One-sample t-tests in R

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

##
##  One Sample t-test
##
## data:  x
## t = -1.3997, df = 99, p-value = 0.1647
## alternative hypothesis: true mean is not equal to 10
## 95 percent confidence interval:
##   9.723872 10.047693
## sample estimates:
## mean of x
##  9.885782
```

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] -1.399742
```

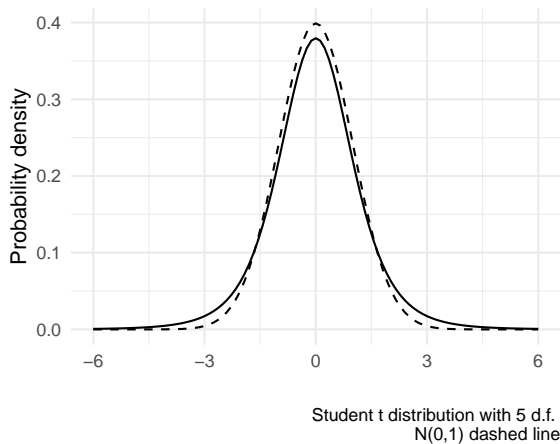
```
p <- pt(abs(round(t,5)), df = N-1, lower.tail = FALSE)  
print(round(p*2,4)) # p*2 = two-tailed p-value
```

```
## [1] 0.1647
```

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

Statistics review

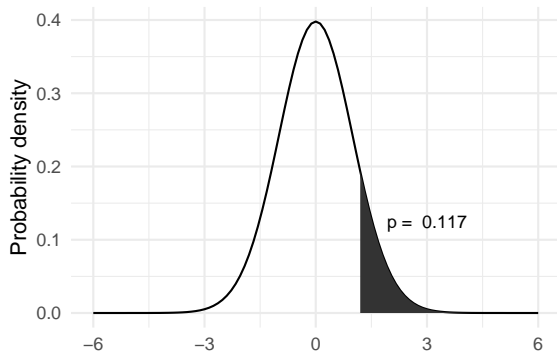
The t-test and the Student t distribution



Statistics review

Area under one tail

$t = 1.2$

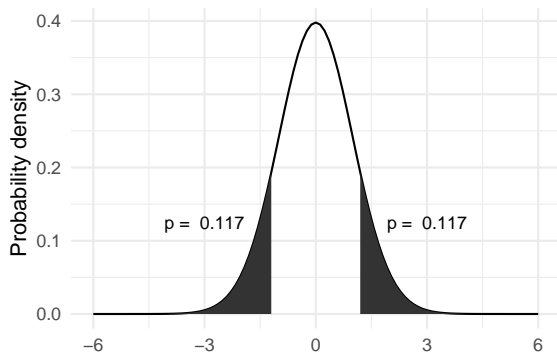


Student t distribution with 99 degrees of freedom.

Statistics review

Area under both tails

$t = 1.2$

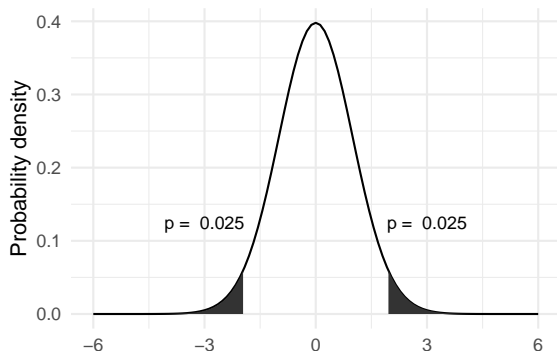


Student t distribution with 99 degrees of freedom.

Statistics review

Area under both tails

$t = 1.96$



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 a set of 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
```

If we repeatedly sample, 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

```
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] 22.343
```

Statistics review

Testing for differences in means

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

```
##
```

```
##  Welch Two Sample t-test
```

```
##
```

```
## data:  x and y
```

```
## t = 22.343, df = 1997.5, p-value < 2.2e-16
```

```
## alternative hypothesis: true difference in means is not
```

```
## 95 percent confidence interval:
```

```
##  0.9121828 1.0877258
```

```
## sample estimates:
```

```
## mean of x mean of y
```

```
##  11.99305  10.99309
```


Statistics review

Type I and Type II errors

- ▶ A Type I error occurs when we incorrectly reject the null hypothesis (“false positive”)
 - ▶ e.g. In the one-sample test, we reject the hypothesis that $\bar{x} = 10$ given $\mu = 10$
- ▶ A Type II error occurs when we incorrectly fail to reject the null hypothesis (“false negative”)
 - ▶ 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

Statistics review

Sign and magnitude errors

- ▶ Gelman, Hill and Vehtari draw attention to two kinds of errors that are often overlooked:
 - ▶ 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 $x = 10y$ but $x = 0.1y$

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

$$\text{var}(x + y) = \text{var}(x) + \text{var}(y) + 2\text{cov}(x, y)$$

- ▶ If x and y are independent then $\text{cov}(x, y) = 0$ and $\text{var}(x + y) = \text{var}(x) + \text{var}(y)$

Statistics review

Covariance

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

$$\text{cov}(x, y) = E[xy] - E[x]E[y]$$

- ▶ For a population, the covariance is

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

- ▶ Sample covariance is defined as

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

Statistics review

Correlation

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

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

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

Why R?

- ▶ Free and open-source
- ▶ Multiple uses
 - ▶ A statistical programming language
 - ▶ Many cutting-edge approaches now implemented in R before Stata
 - ▶ Alongside Python, it is one of the main programming languages used by data scientists
 - ▶ Unfortunately, Stata does not have such flexibility
- ▶ A very active developer community
 - ▶ Lots of cool 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
 - ▶ Terminal interface
 - ▶ Integrations including Github and Python

RMarkdown

Overview

- ▶ RMarkdown is an interactive coding environment
 - ▶ RMarkdown documents can combine text, LaTeX code, R code, and any output.
 - ▶ Write in Markdown or Visual Editor
 - ▶ 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