# 6. Hypothesis testing and confidence intervals

We have the following goals:

- Understand how to construct confidence intervals for parameters in a linear model.
- Understand how to test statistical hypotheses about a linear model.
- In particular, to ask and answer the question: "Are the data consistent with a hypothesis that a covariate, or a collection of covariates, are unimportant?" (What is the fundamental scientific importance of the slightly contorted logical reasoning in this question?)
- Learn to use R to carry out these tasks.
- See how the linear model includes and extends basic tests for means of one and two samples.

# Confidence intervals

- An interval $[u, v]$ constructed using the data $\mathbf{y}$ is said to **cover** a parameter $\theta$ if $u \leq \theta \leq v$.
- $[u, v]$ is a 95% **confidence interval** (CI) for $\theta$ if the same construction, applied to a large number of draws from the model, would cover $\theta$ 95% of the time.
- A **parameter** is a name for any unknown constant in a model. In linear models,each component $\beta_1, \ldots, \beta_p$ of the **coefficient vector** $\boldsymbol{\beta}$ is a parameter. So is the variance $\sigma^2$ of the measurement error.
- A confidence interval is the usual way to represent the amount of uncertainty in an estimated parameter.
- The parameter is not random. According to the model, it has a fixed but unknown value.The observed interval $[u, v]$ is also not random. An interval $[U, V]$ constructed using a vector of random variables $\mathbf{Y}$ defined in a probability model is random.
- If the model is appropriate, then it is reasonable to treat the data $\mathbf{y}$ like a realization from the probability model.

# A confidence interval for the coefficient of a linear model

- Consider estimating $\beta_1$ in the linear model $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.
- Recall that $\mathrm{E}[\hat{\beta}_1] = \beta_1$ and $\mathrm{SD}(\hat{\beta}_1) = \sigma \sqrt{\left[\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\right]_{11}}$.

**Question 6.1**. Supposing we can make a normal approximation, show that $\mathrm{P}\big[\, \hat{\beta}_1 - 1.96\,\mathrm{SD}(b_1) \le \beta_1 \le \hat{\beta}_1 + 1.96\,\mathrm{SD}(b_1) \,\big] = 0.95$

- Therefore, an approximate 95% CI for $\beta_1$ is

$$\boxed{\big[\, b_1 - 1.96\,\mathrm{SE}(b_1)\,,\; b_1 + 1.96\,\mathrm{SE}(b_1) \,\big]}$$

where $\mathbf{y} = \mathbb{X}\mathbf{b} + \mathbf{e}$ with $\mathrm{SE}(b_1) = s \sqrt{\left[\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\right]_{11}}$.

# A CI for association between unemployment and mortality

```
c1 <- summary(lm(L_detrended~U_detrended))$coefficients ; c1


##                Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 0.2899928 0.09343146 3.103802 0.002812739
## U_detrended 0.1313673 0.06321939 2.077959 0.041606370


beta_U <- c1["U_detrended","Estimate"]
SE_U <- c1["U_detrended","Std. Error"]
z <- qnorm(1-0.05/2) # for a 95% CI using a normal approximation
cat("CI = [", beta_U - z * SE_U, ",", beta_U + z * SE_U, "]")


## CI = [ 0.0074596 , 0.2552751 ]
```

**Interpretation**. We appear to have found evidence that each percentage point of unemployment above trend is associated with about 0.13 years of additional life expectancy. The 95% CI doesn't include zero.

**Question 6.2**. Do you believe this discovery? How could you criticize it?

# Association is not causation

"Whatever phenomenon varies in any manner whenever another phenomenon varies in some particular manner, is either a cause or an effect of that phenomenon, or is connected with it through some fact of causation." *(John Stuart Mill, A System of Logic, Vol. 1. 1843. p. 470.*

- Put differently: If $A$ and $B$ are associated statistically, we can infer that either $A$ causes $B$, or $B$ causes $A$, or both have some common cause $C$.
- A useful mantra: **Association is not causation.**
- Writing a linear model where $A$ depends on $B$ can show association but we need extra work to argue for causation. We need to rule out $B$ causing $A$ and the possibility of any common cause $C$.

**Question 6.3**. Discuss the extent to which the association between detrended unemployment and life expectancy can and cannot be interpreted causally.

# A review of progress so far in this course

**Producing and understanding this confidence interval for a linear model brought together all the things we've done so far in this course.**

- We needed to get the data into a computer and run statistical software.
- To understand what the computer was doing for us, and help us to command it correctly, we needed to know about:
  1. matrices
  2. writing a linear model and fitting it by least squares
  3. probability models
  4. expectation and variance
  5. the normal distribution

**You could run computer code by learning to follow line-by-line instructions without understanding what the instructions do. But then you wouldn't be in control of your own data analysis.**

# Hypothesis tests

- We try to see patterns in our data. We hope to discover phenomena that will advance science, or help the environment, or reduce sickness and poverty, or make us rich, ...
- How can we tell whether our new theory is like seeing animals or faces in the clouds?
- From Wikipedia: "**Pareidolia** is a psychological phenomenon in which the mind responds to a stimulus ... by perceiving a familiar pattern where none exists (e.g. in random data)".
- The research community has set a standard: The evidence presented to support a new theory should be unlikely under a **null hypothesis** that the new theory is false. To quantify *unlikely* we need a probability model.

# Hypothesis tests and the scientific method

- From a different perspective, a standard view of scientific progress holds that scientific theories cannot be proved correct, they can only be falsified (`https://en.wikipedia.org/wiki/Falsifiability`).
- Accordingly, scientists look for evidence to refute the **null hypothesis** that data can be explained by current scientific understanding.
- If the null hypothesis is inadequate to explain data, the scientist may propose an **alternative hypothesis** which better explains these data.
- The alternative hypothesis will subsequently be challenged with new data.

# The scientific method in statistical language

1. **Ask a question**
2. **Obtain relevant data**.
3. **Write a null and alternative hypothesis to represent your question in a probability model**. This may involve writing a linear model so that $\beta_1 = 0$ corresponds to the null hypothesis of "no effect" and $\beta_1 \neq 0$ is a discovered "effect."
4. **Calculate a test statistic**. The test statistic is a quantity computed using the data that summarizes the evidence against the null hypothesis. For our linear model example, the least squares coefficient $b_1$ is a natural statistics to test the hypothesis $\beta_1 = 0$.
5. **Calculate the p-value**, which is the probability that the model generates a test statistic at least as extreme as that observed. For our linear model example, the p-value is $\mathrm{P}\left[|\hat{\beta}_1| > |b_1|\right]$. We can find this probability, when $\beta_1 = 0$, using a normal approximation.
6. **Conclusions**. A small p-value (often, $< 0.05$) is evidence for **rejecting** the null hypothesis. The data analysis may suggest new questions: **Return to Step 1**.

# Using confidence intervals to construct a hypothesis test

- It is often convenient to use the confidence interval as a test statistic.
- If the confidence interval doesn't cover the null hypothesis, then we have evidence to reject that null hypothesis.
- If we do this test using a 95% confidence interval, we have a 5% chance that we reject the null hypothesis if it is true. This follows from the definition of a confidence interval: whatever the true unknown value of the parameter, the confidence interval covers it with probability 0.95.

# Some notation for hypothesis tests

- The null hypothesis is $H_0$ and the alternative is $H_a$.
- We write $t$ for the test statistic calculated using the data $\mathbf{y}$. We write $T$ for the random variable constructed by calculating the test statistic using a random vector $\mathbf{Y}$ drawn from the probability model under $H_0$.
- The p-value is $\text{pval} = \text{P}\big[|T| \geq |t|\big]$. Here, we are assuming "extreme" means "large in magnitude." Occasionally, it may make more sense to use $\text{pval} = \text{P}\big[T \geq t\big]$.
- We reject $H_0$ at **significance level** $\alpha$ if $\text{pval} < \alpha$. Common choices of $\alpha$ are $\alpha = 0.05$, $\alpha = 0.01$, $\alpha = 0.001$.

**Question 6.4**. When we report the results of a hypothesis test, we can either (i) give the p-value, or (ii) say whether $H_0$ is rejected at a particular significance level. What are the advantages and disadvantages of each?

# A hypothesis test for unemployment and mortality

**Question 6.5**. Write a formal hypothesis test of the null hypothesis that there is no association between unemployment and mortality. Compute a p-value using a normal approximation. What do you think is an appropriate significance level $\alpha$ for deciding whether to reject the null hypothesis?

# Normal approximations versus Student's t distribution

- Notice that 'summary(lm(...))' gives 't value' and 'Pr(¿—t—)'.
- The 't value' is the estimated coefficient divided by its standard error. This measures how many standard error units the estimated coefficient is from zero.
- 'Pr(¿—t—)' is similar, but slightly larger, than the p-value coming from the normal approximation.
- R is using Student's t distribution, which makes allowance for chance variation from using $s$ as an approximation to $\sigma$ when we compute the standard error.
- R uses a t random variable to model the distribution of the statistic $t$. Giving the full name (Student's t distribution) may add clarity.
- With sophisticated statistical methods, it is often hard to see if they work well just by reading about them. Fortunately, it is often relatively easy to do a **simulation study** to see what is going on.

# Simulating from Student's t distribution

- Suppose $X, X_1, \ldots, X_d$ are $d+1$ independent identially distributed (iid) normal random variables with mean zero and standard deviation $\sigma$.
- We write $X, X_1, \ldots, X_d \sim \text{iid } N[0, \sigma]$.
- Student's t distribution on $d$ degrees of freedom is defined to be the distribution of $T = X/\hat{\sigma}$ where $\hat{\sigma} = \sqrt{\frac{1}{d}\sum_{i=1}^{d} X_i^2}$.
- A normal approximation would say $T$ is approximately $N[0,1]$ since $\hat{\sigma}$ is an estiamte of $\sigma$.
- With a computer, we can simulate $T$ many times, plot a histogram, and compare it to the probability density function of the normal distribution and Student's t distribution.
- The goals in doing this:
  1. Some practice working with Student's t distribution.
  2. Finding how the t distribution compares to the normal distribution as $d$ varies.
  3. Practice the skill of designing a simulation experiment.

- Let's start by simulating a matrix `X` of iid normal random variables.

```
N <- 1000 ; sigma <- 1 ; d <- 10
X <- matrix(rnorm(N*(d+1),mean=0,sd=sigma),nrow=N)
```

- Now, we write a function that computes $T$ given $X_1, \ldots, X_d, X$

```
T_evaluator <- function(x) x[d+1] / sqrt(sum(x[1:d]^2)/d)
```

- Then, use `apply()` to evaluate $T$ on each row of 'X'.

```
T_simulated <- apply(X,1,T_evaluator)
```

- A histogram of these simulations can be compared with the normal density and the t density

```
hist(T_simulated)
```