

# Introduction<sup>1</sup>

STA442/2101 Fall 2014

---

<sup>1</sup>See last slide for copyright information.

# Background Reading

Optional

- ▶ Chapter 1 of *Linear models with R*
- ▶ Chapter 1 of Davison's *Statistical models: Data, and probability models for data*.

## Goal of statistical analysis

The goal of statistical analysis is to draw reasonable conclusions from noisy numerical data.

# Steps in the process of statistical analysis

## One approach

- ▶ Consider a fairly realistic example or problem.
- ▶ Decide on a statistical model.
- ▶ Perhaps decide sample size.
- ▶ Acquire data.
- ▶ Examine and clean the data; generate displays and descriptive statistics.
- ▶ Estimate model parameters, for example by maximum likelihood.
- ▶ Carry out tests, compute confidence intervals, or both.
- ▶ Perhaps re-consider the model and go back to estimation.
- ▶ Based on the results of estimation and inference, draw conclusions about the example or problem.

# What is a statistical model?

You should always be able to state the model.

A *statistical model* is a set of assertions that partly specify the probability distribution of the observable data. The specification may be direct or indirect.

- ▶ Let  $X_1, \dots, X_n$  be a random sample from a normal distribution with expected value  $\mu$  and variance  $\sigma^2$ .
- ▶ For  $i = 1, \dots, n$ , let  $Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \epsilon_i$ , where

$\beta_0, \dots, \beta_k$  are unknown constants.

$x_{ij}$  are known constants.

$\epsilon_1, \dots, \epsilon_n$  are independent  $N(0, \sigma^2)$  random variables.

$\sigma^2$  is an unknown constant.

$Y_1, \dots, Y_n$  are observable random variables.

Is the model the same thing as the *truth*?

## Parameter Space

The *parameter space* is the set of values that can be taken on by the parameter.

- ▶ Let  $X_1, \dots, X_n$  be a random sample from a normal distribution with expected value  $\mu$  and variance  $\sigma^2$ .  
The parameter space is  $\{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$ .
- ▶ For  $i = 1, \dots, n$ , let  $Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \epsilon_i$ , where
  - $\beta_0, \dots, \beta_k$  are unknown constants.
  - $x_{ij}$  are known constants.
  - $\epsilon_1, \dots, \epsilon_n$  are independent  $N(0, \sigma^2)$  random variables.
  - $\sigma^2$  is an unknown constant.
  - $Y_1, \dots, Y_n$  are observable random variables.

The parameter space is

$$\{(\beta_0, \dots, \beta_k, \sigma^2) : -\infty < \beta_j < \infty, \sigma^2 > 0\}.$$

## Coffee taste test

A fast food chain is considering a change in the blend of coffee beans they use to make their coffee. To determine whether their customers prefer the new blend, the company plans to select a random sample of  $n = 100$  coffee-drinking customers and ask them to taste coffee made with the new blend and with the old blend, in cups marked “A” and “B.” Half the time the new blend will be in cup A, and half the time it will be in cup B. Management wants to know if there is a difference in preference for the two blends.

## Statistical model

Letting  $\theta$  denote the probability that a consumer will choose the new blend, treat the data  $X_1, \dots, X_n$  as a random sample from a Bernoulli distribution. That is, independently for  $i = 1, \dots, n$ ,

$$P(x_i|\theta) = \theta^{x_i}(1 - \theta)^{1-x_i}$$

for  $x_i = 0$  or  $x_i = 1$ , and zero otherwise.

- ▶ Parameter space is the interval from zero to one.
- ▶  $\theta$  could be estimated by maximum likelihood.
- ▶ Large-sample tests and confidence intervals are available.



## Tests of statistical hypotheses

- ▶ Model:  $X \sim F_\theta$
- ▶  $X$  is the data vector, and  $\mathcal{X}$  is the sample space:  $X \in \mathcal{X}$
- ▶  $\theta$  is the parameter, and  $\Theta$  is the parameter space:  $\theta \in \Theta$
- ▶ Null hypothesis is  $H_0 : \theta \in \Theta_0$  v.s.  $H_A : \theta \in \Theta \cap \Theta_0^c$
- ▶ Meaning of the *null* hypothesis is that *nothing* interesting is happening.
- ▶  $\mathcal{C} \subset \mathcal{X}$  is the *critical region*. Reject  $H_0$  in favour of  $H_A$  when  $X \in \mathcal{C}$ .
- ▶ Significance level  $\alpha$  (*size* of the test) is the maximum probability of rejecting  $H_0$  when  $H_0$  is true.
- ▶  $p$ -value is the smallest value of  $\alpha$  for which  $H_0$  can be rejected.
- ▶ Small  $p$ -values are interpreted as providing stronger evidence against the null hypothesis.

# Type I and Type II error

A Neyman-Pearson idea rather than Fisher

- ▶ Type I error is to reject  $H_0$  when  $H_0$  is true.
- ▶ Type II error is to *not* reject  $H_0$  when  $H_0$  is false.
- ▶  $1 - Pr\{\text{Type II Error}\}$  is called *power*.
- ▶ Power may be used to select sample size.

## Carry out a test to determine which brand of coffee is preferred

Recall the model is  $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} B(1, \theta)$

Start by stating the null hypothesis.

- ▶  $H_0 : \theta = 0.50$
- ▶  $H_1 : \theta \neq 0.50$
- ▶ Could you make a case for a one-sided test?
- ▶  $\alpha = 0.05$  as usual.
- ▶ Central Limit Theorem says  $\hat{\theta} = \bar{X}$  is approximately normal with mean  $\theta$  and variance  $\frac{\theta(1-\theta)}{n}$ .

Several valid test statistics for  $H_0 : \theta = \theta_0$  are available

Two of them are

$$Z_1 = \frac{\sqrt{n}(\bar{X} - \theta_0)}{\sqrt{\theta_0(1 - \theta_0)}}$$

and

$$Z_2 = \frac{\sqrt{n}(\bar{X} - \theta_0)}{\sqrt{\bar{X}(1 - \bar{Y})}}$$

What is the critical value? Your answer is a number.

```
> alpha = 0.05
> qnorm(1-alpha/2)
[1] 1.959964
```

## Calculate the test statistic and the $p$ -value for each test

Suppose 60 out of 100 preferred the new blend

```
> theta0 = .5; ybar = .6; n = 100
> Z1 = sqrt(n)*(ybar-theta0)/sqrt(theta0*(1-theta0)); Z1
[1] 2
> pval1 = 2 * (1-pnorm(Z1)); pval1
[1] 0.04550026

> Z2 = sqrt(n)*(ybar-theta0)/sqrt(ybar*(1-ybar)); Z2
[1] 2.041241
> pval2 = 2 * (1-pnorm(Z2)); pval2
[1] 0.04122683
```

## Conclusions

- ▶ Do you reject  $H_0$ ? *Yes, just barely.*
- ▶ Isn't the  $\alpha = 0.05$  significance level pretty arbitrary? *Yes, but if people insist on a Yes or No answer, this is what you give them.*
- ▶ What do you conclude, in symbols?  $\theta \neq 0.50$ . *Specifically,  $\theta > 0.50$ .*
- ▶ What do you conclude, in plain language? Your answer is a statement about coffee. *More consumers prefer the new blend of coffee beans.*
- ▶ Can you really draw directional conclusions when all you did was reject a non-directional null hypothesis? *Yes. Decompose the two-sided size  $\alpha$  test into two one-sided tests of size  $\alpha/2$ . This approach works in general.*

It is very important to state directional conclusions, and state them clearly in terms of the subject matter. **Say what happened!** If you are asked state the conclusion in plain language, your answer *must* be free of statistical mumbo-jumbo.

## What about negative conclusions?

What would you say if  $Z = 1.84$ ?

Here are two possibilities, in plain language.

- ▶ “This study does not provide clear evidence that consumers prefer one blend of coffee beans over the other.”
- ▶ “The results are consistent with no difference in preference for the two coffee bean blends.”

In this course, we will not just casually *accept* the null hypothesis.

# Confidence intervals

Usually for individual parameters

- ▶ Point estimates may give a false sense of precision.
- ▶ We should provide a margin of probable error as well.



Give a 95% confidence interval for the taste test data.

The answer is a pair of numbers. Show some work.

$$\begin{aligned} & \left( \bar{x} - z_{\alpha/2} \sqrt{\frac{\bar{x}(1-\bar{x})}{n}} , \bar{x} + z_{\alpha/2} \sqrt{\frac{\bar{x}(1-\bar{x})}{n}} \right) \\ &= \left( 0.60 - 1.96 \sqrt{\frac{0.6 \times 0.4}{100}} , 0.60 + 1.96 \sqrt{\frac{0.6 \times 0.4}{100}} \right) \\ &= (0.504, 0.696) \end{aligned}$$

In a report, you could say

- ▶ The estimated proportion preferring the new coffee bean blend is  $0.60 \pm 0.096$ , or
- ▶ “Sixty percent of consumers preferred the new blend. These results are expected to be accurate within 10 percentage points, 19 times out of 20.”

## Meaning of the confidence interval

- ▶ We calculated a 95% confidence interval of (0.504, 0.696) for  $\theta$ .
- ▶ Does this mean  $Pr\{0.504 < \theta < 0.696\} = 0.95$ ?
- ▶ No! The quantities 0.504, 0.696 and  $\theta$  are all constants, so  $Pr\{0.504 < \theta < 0.696\}$  is either zero or one.
- ▶ The endpoints of the confidence interval are random variables, and the numbers 0.504 and 0.696 are *realizations* of those random variables, arising from a particular random sample.
- ▶ Meaning of the probability statement: If we were to calculate an interval in this manner for a large number of random samples, the interval would contain the true parameter around 95% of the time.
- ▶ The confidence interval is a guess, and the guess is either right or wrong. But the guess is the constructed by a method that is right 95% of the time.

## More on confidence intervals

- ▶ Can have confidence *regions* for the entire parameter vector or multi-dimensional functions of the parameter vector.
- ▶ Confidence regions correspond to tests.

## Copyright Information

This slide show was prepared by **Jerry Brunner**, Department of Statistics, University of Toronto. It is licensed under a **Creative Commons Attribution - ShareAlike 3.0 Unported License**. Use any part of it as you like and share the result freely. The  $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$  source code is available from the course website:  
<http://www.utstat.toronto.edu/~brunner/oldclass/appliedf14>