Methods of Applied Statistics I[1]

STA2101 Fall 2019

[1] See last slide for copyright information.

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 $X_1, \ldots, X_n$ be a random sample from a normal distribution with expected value μ and variance σ^2. The parameters μ and σ^2 are unknown.
- For $i = 1, \ldots, n$, let $y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i$, where

  $\beta_0, \ldots, \beta_{p-1}$ are unknown constants.
  $x_{i,j}$ are known constants.
  $\epsilon_1, \ldots, \epsilon_n$ are independent $N(0, \sigma^2)$ random variables.
  σ^2 is an unknown constant.
  $y_1, \ldots, y_n$ are observable random variables.

  The parameters $\beta_0, \ldots, \beta_{p-1}, \sigma^2$ are unknown.

Model and Truth

Is a statistical model the same thing as the truth?

"Essentially all models are wrong, but some are useful." (Box and Draper, 1987, p. 424)

Parameter Space

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

- Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with expected value μ and variance σ^2.
  The parameter space is $\{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$.
- For $i = 1, \ldots, n$, let $y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i$, where

  $\beta_0, \ldots, \beta_{p-1}$ are unknown constants.
  $x_{i,j}$ are known constants.
  $\epsilon_1, \ldots, \epsilon_n$ are independent $N(0, \sigma^2)$ random variables.
  σ^2 is an unknown constant.
  $y_1, \ldots, y_n$ are observable random variables.

  The parameter space is
  $\{(\beta_0, \ldots, \beta_{p-1}, \sigma^2) : -\infty < \beta_j < \infty, \sigma^2 > 0\}$.

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 θ denote the probability that a consumer will choose the new blend, treat the data $Y_1, \ldots, Y_n$ as a random sample from a Bernoulli distribution. That is, independently for $i = 1, \ldots, n$,

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

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

- Parameter space is the interval from zero to one.
- θ could be estimated by maximum likelihood. $\widehat{\theta} = \overline{y}$.
- Large-sample tests and confidence intervals are available.

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

Recall the model is $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} B(1, \theta)$

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 $\widehat{\theta} = \overline{Y}$ is approximately normal with mean θ and variance $\frac{\theta(1-\theta)}{n}$.

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

Recall that approximately, $\overline{Y} \sim N(\theta, \frac{\theta(1-\theta)}{n})$

Two of them are

$$Z_1 = \frac{\sqrt{n}(\overline{Y} - \theta_0)}{\sqrt{\theta_0(1-\theta_0)}}$$

and

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

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

$$Z_1 = \frac{\sqrt{n}(\overline{Y} - \theta_0)}{\sqrt{\theta_0(1-\theta_0)}}$$

```
> 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
```

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

```
> 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.*

A technical issue

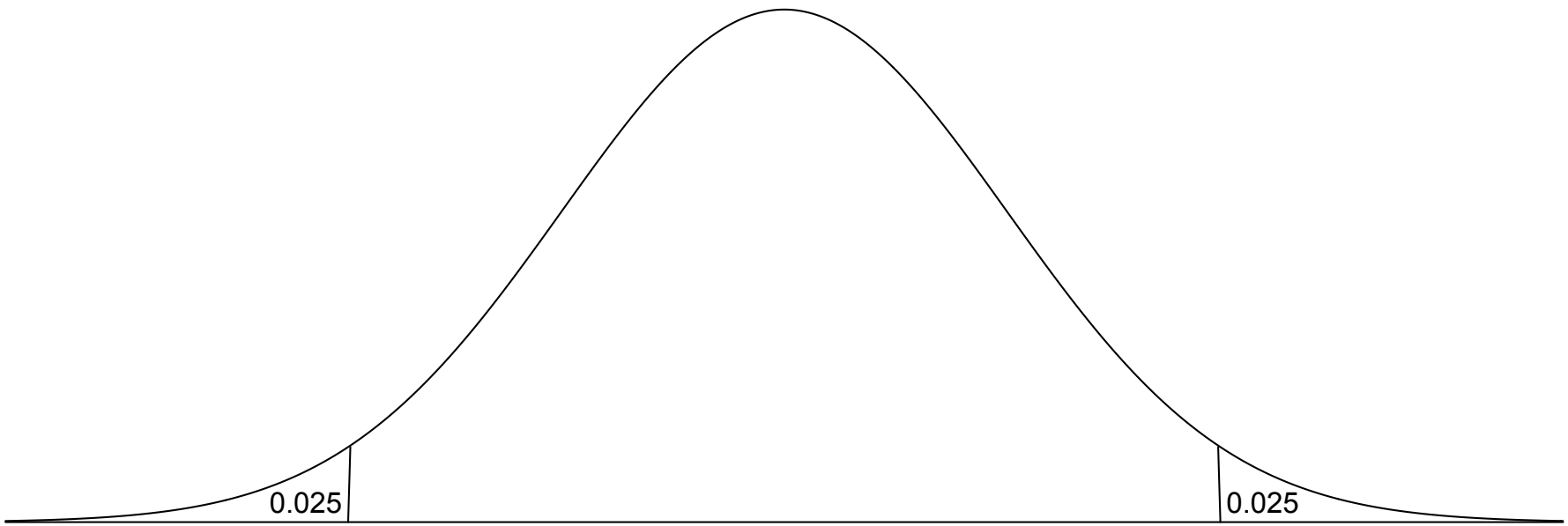
- In this class we will mostly avoid one-tailed tests.
- Why? Ask what would happen if the results were strong and in the opposite direction to what was predicted (dental example).
- But when H_0 is rejected, we still draw directional conclusions.
- For example, if x is income and y is credit card debt, we test $H_0 : \beta_1 = 0$ with a two-sided t-test.
- Say $p = 0.0021$ and $\widehat{\beta}_1 = 1.27$. We say "Consumers with higher incomes tend to have more credit card debt."
- Is this justified? We'd better hope so, or all we can say is "There is a connection between income and average credit card debt."
- Then they ask: "What's the connection? Do people with lower income have more debt?"
- And you have to say "Sorry, I don't know."

The technical resolution

Decompose the two-sided test into a set of two one-sided tests with significance level $\alpha/2$, equivalent to the two-sided test.

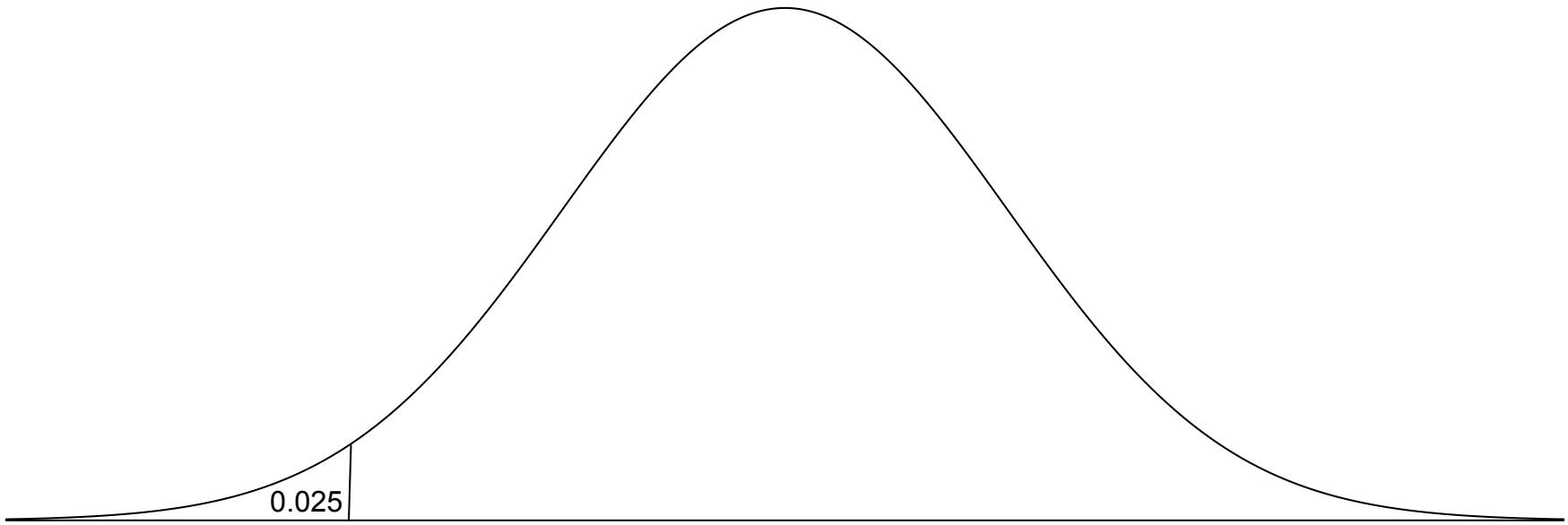
Two-sided test

$$H_0 : \theta = \tfrac{1}{2} \text{ versus } H_1 : \theta \neq \tfrac{1}{2},\ \alpha = 0.05$$



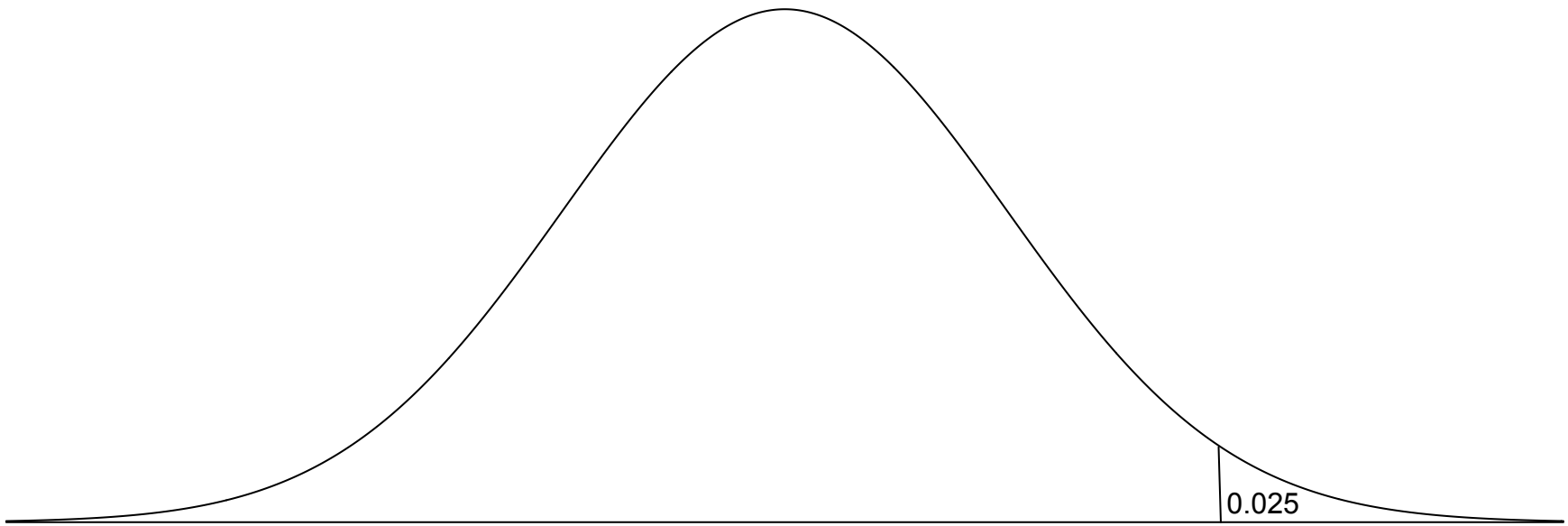
Left-sided test

$$H_0 : \theta \geq \tfrac{1}{2} \text{ versus } H_1 : \theta < \tfrac{1}{2},\ \alpha = 0.05$$



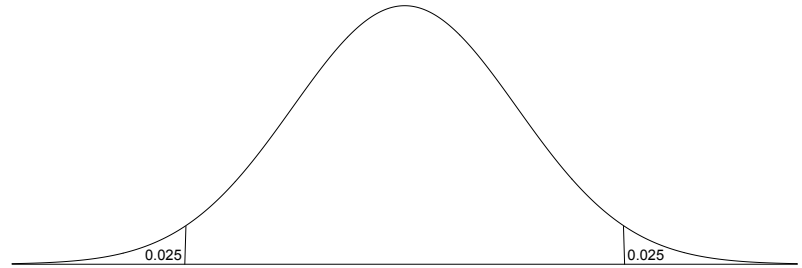
Right-sided test

$$H_0 : \theta \leq \tfrac{1}{2} \text{ versus } H_1 : \theta > \tfrac{1}{2},\ \alpha = 0.05$$

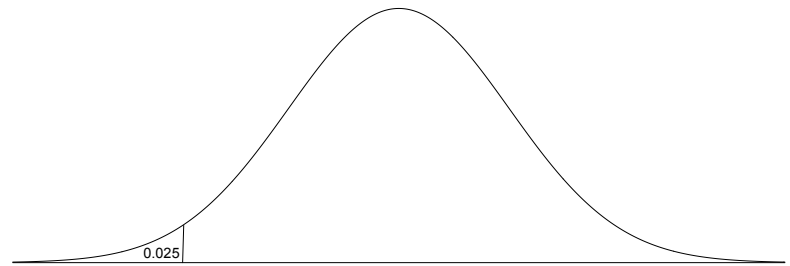


Decomposing the 2-sided test into two 1-sided tests

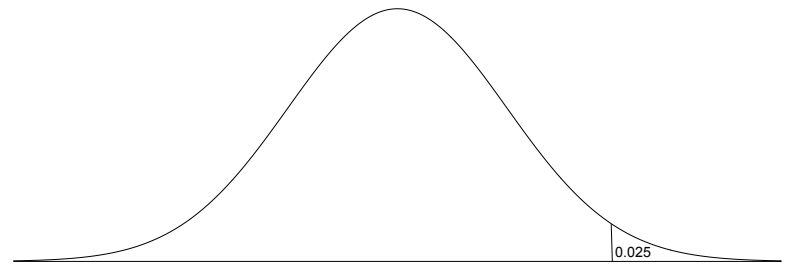
$H_0 : \theta = \frac{1}{2}$ vs. $H_1 : \theta \neq \frac{1}{2}$, $\alpha = 0.05$



$H_0 : \theta \geq \frac{1}{2}$ vs. $H_1 : \theta < \frac{1}{2}$, $\alpha = 0.05$



$H_0 : \theta \leq \frac{1}{2}$ versus $H_1 : \theta > \frac{1}{2}$, $\alpha = 0.05$



- Clearly, the 2-sided test rejects H_0 if and only if exactly *one* of the 1-sided tests reject H_0.
- Carry out *both* of the one-sided tests.
- Draw a directional conclusion if H_0 is rejected.

Summary of the technical resolution

- Decompose the two-sided test into a set of two one-sided tests with significance level $\alpha/2$, equivalent to the two-sided test.
- In practice, just look at the sign of the regression coefficient, or compare the sample means.
- Under the surface you are decomposing the two-sided test, but you never mention it.

Plain language

- 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.
- *Marking rule*: If the question asks for plain language and you draw a non-directional conclusion when a directional conclusion is possible, you get half marks at most.

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. We will *not* say that there was no difference in preference.

We are taking the side of Fisher over Neyman and Pearson in an old and very nasty argument.

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.

Copyright Information