A Frequentist Introduction[1]

STA442/2101 Fall 2018

[1]See last slide for copyright information.

Background Reading

Optional

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 $X_1, \ldots, X_n$ 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.
- Large-sample tests and confidence intervals are available.

Note that $Y = \sum_{i=1}^{n} Y_i$ is the number of consumers who choose the new blend. Because $Y \sim B(n, \theta)$, the whole experiment could also be treated as a single observation from a Binomial.

Find the MLE of θ

Show your work

Denoting the likelihood by $L(\theta)$ and the log likelihood by $\ell(\theta) = \log L(\theta)$, maximize the log likelihood.

$$
\begin{aligned}
\frac{\partial \ell}{\partial \theta} &= \frac{\partial}{\partial \theta} \log \left(\prod_{i=1}^{n} P(y_i|\theta) \right) \\
&= \frac{\partial}{\partial \theta} \log \left(\prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} \right) \\
&= \frac{\partial}{\partial \theta} \log \left(\theta^{\sum_{i=1}^{n} y_i} (1-\theta)^{n-\sum_{i=1}^{n} y_i} \right) \\
&= \frac{\partial}{\partial \theta} \left((\sum_{i=1}^{n} y_i) \log \theta + (n - \sum_{i=1}^{n} y_i) \log(1-\theta) \right) \\
&= \frac{\sum_{i=1}^{n} y_i}{\theta} - \frac{n - \sum_{i=1}^{n} y_i}{1-\theta}
\end{aligned}
$$

Setting the derivative to zero and solving

- $\theta = \frac{\sum_{i=1}^{n} y_i}{n} = \overline{y}$
- Second derivative test: $\frac{\partial^2 \log \ell}{\partial \theta^2} = -n \left(\frac{1-\overline{y}}{(1-\theta)^2} + \frac{\overline{y}}{\theta^2} \right) < 0$
- Concave down, maximum, and the MLE is the sample proportion: $\widehat{\theta} = \overline{y} = p$

Numerical estimate

Suppose 60 of the 100 consumers prefer the new blend. Give a point estimate the parameter θ. Your answer is a number.

```
> p = 60/100; p
[1] 0.6
```

Tests of statistical hypotheses

- Model: $Y \sim F_\theta$
- Y is the data vector, and $\mathcal{Y}$ is the sample space: $Y \in \mathcal{Y}$
- θ is the parameter, and Θ 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.
- $\mathcal{C} \subset \mathcal{Y}$ is the *critical region*. Reject H_0 in favour of H_A when $Y \in \mathcal{C}$.
- Significance level α (*size* of the test) is the maximum probability of rejecting H_0 when H_0 is true. Conventionally, $\alpha = 0.05$.
- p-value is the smallest value of α 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*.
- If two tests have the same maximum Type I error probability α, the one with higher power is better.
- Power may also be used to select sample size.

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

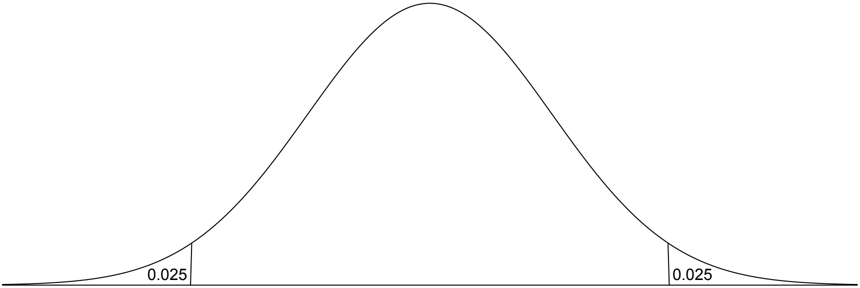
- In this class we will mostly avoid one-tailed tests.
- Why? Ask what would happen if the results were strong and in the opposite direction to what was predicted (dental example).
- But when H_0 is rejected, we still draw directional conclusions.
- For example, if x is income and y is credit card debt, 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."
- It's a good way to get fired, or at least look silly.

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.

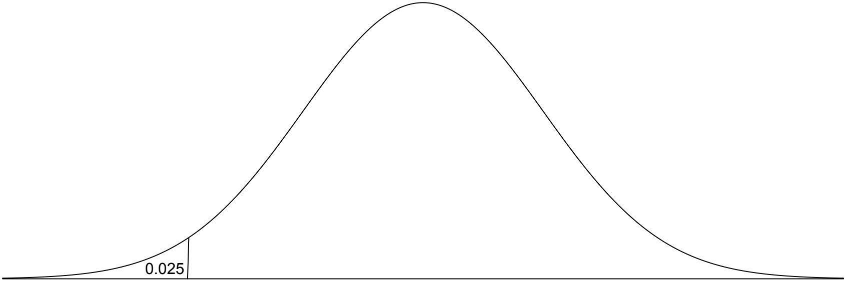
Two-sided test

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



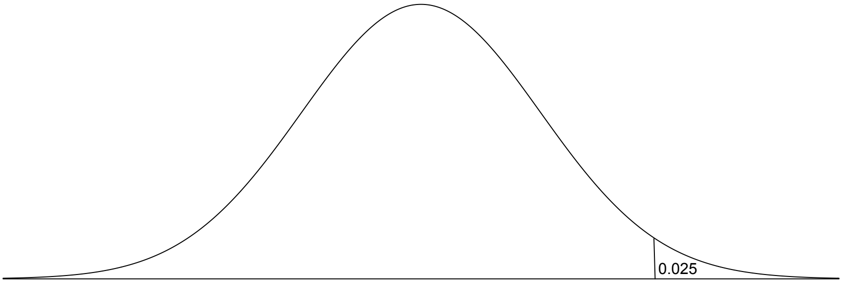
Left-sided test

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



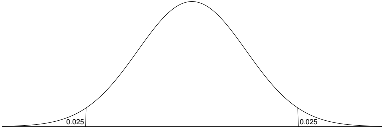
Right-sided test

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

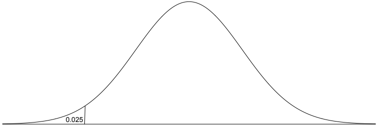


Decomposing the 2-sided test into two 1-sided tests

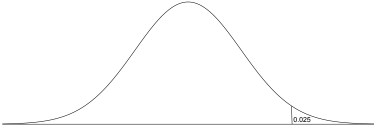
$H_0 : \theta = \frac{1}{2}$ vs. $H_1 : \theta \neq \frac{1}{2}$, $\alpha = 0.05$



$H_0 : \theta \geq \frac{1}{2}$ 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 philosophic dispute.

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.

Confidence Intervals

Taste test example

Approximately for large n,

$$
\begin{aligned}
1-\alpha &= Pr\{-z_{\alpha/2} < Z < z_{\alpha/2}\} \\
&\approx Pr\left\{-z_{\alpha/2} < \frac{\sqrt{n}(\overline{Y}-\theta)}{\sqrt{\overline{Y}(1-\overline{Y})}} < z_{\alpha/2}\right\} \\
&= Pr\left\{\overline{Y} - z_{\alpha/2}\sqrt{\frac{\overline{Y}(1-\overline{Y})}{n}} < \theta < \overline{Y} + z_{\alpha/2}\sqrt{\frac{\overline{Y}(1-\overline{Y})}{n}}\right\}
\end{aligned}
$$

- Could express this as $\overline{Y} \pm z_{\alpha/2}\sqrt{\frac{\overline{Y}(1-\overline{Y})}{n}}$.
- $z_{\alpha/2}\sqrt{\frac{\overline{Y}(1-\overline{Y})}{n}}$ is sometimes called the *margin of error*.
- If $\alpha = 0.05$, it's the 95% margin of error.

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

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

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

In a report, you could say

- The estimated proportion preferring the new coffee bean blend is 0.60 ± 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 θ.
- Does this mean $Pr\{0.504 < \theta < 0.696\} = 0.95$?
- No! The quantities 0.504, 0.696 and θ 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.

Confidence intervals (regions) correspond to tests

Recall $Z_1 = \frac{\sqrt{n}(\overline{Y}-\theta_0)}{\sqrt{\theta_0(1-\theta_0)}}$ and $Z_2 = \frac{\sqrt{n}(\overline{Y}-\theta_0)}{\sqrt{\overline{Y}(1-\overline{Y})}}$.

H_0 is *not* rejected if and only if

$$-z_{\alpha/2} < Z_2 < z_{\alpha/2}$$

if and only if

$$\overline{Y} - z_{\alpha/2}\sqrt{\frac{\overline{Y}(1-\overline{Y})}{n}} < \theta_0 < \overline{Y} + z_{\alpha/2}\sqrt{\frac{\overline{Y}(1-\overline{Y})}{n}}$$

- So the confidence interval consists of those parameter values θ_0 for which $H_0 : \theta = \theta_0$ is *not* rejected.
- That is, the null hypothesis is rejected at significance level α if and only if the value given by the null hypothesis is outside the $(1-\alpha) \times 100\%$ confidence interval.

Selecting sample size

- Where did that $n = 100$ come from?
- Probably off the top of someone's head.
- We can (and should) be more systematic.
- Sample size can be selected
  - To achieve a desired margin of error
  - To achieve a desired statistical power
  - In other reasonable ways

Statistical Power

The power of a test is the probability of rejecting H_0 when H_0 is false.

- More power is good.
- Power is not just one number. It is a *function* of the parameter(s).
- Usually,
  - For any n, the more incorrect H_0 is, the greater the power.
  - For any parameter value satisfying the alternative hypothesis, the larger n is, the greater the power.

Statistical power analysis

To select sample size

- Pick an effect you'd like to be able to detect – a parameter value such that H_0 is false. It should be just over the boundary of interesting and meaningful.
- Pick a desired power, a probability with which you'd like to be able to detect the effect by rejecting the null hypothesis.
- Start with a fairly small n and calculate the power. Increase the sample size until the desired power is reached.

There are two main issues.

- What is an "interesting" or "meaningful" parameter value?
- How do you calculate the probability of rejecting H_0?

Calculating power for the test of a single proportion

True parameter value is θ

$$
\begin{aligned}
\text{Power} &= 1 - Pr\{-z_{\alpha/2} < Z_2 < z_{\alpha/2}\} \\
&= 1 - Pr\left\{-z_{\alpha/2} < \frac{\sqrt{n}(\overline{Y} - \theta_0)}{\sqrt{\overline{Y}(1-\overline{Y})}} < z_{\alpha/2}\right\} \\
&= \ldots \\
&= 1 - Pr\left\{\frac{\sqrt{n}(\theta_0 - \theta)}{\sqrt{\theta(1-\theta)}} - z_{\alpha/2}\sqrt{\frac{\overline{Y}(1-\overline{Y})}{\theta(1-\theta)}} < \frac{\sqrt{n}(\overline{Y} - \theta)}{\sqrt{\theta(1-\theta)}}\right. \\
&\qquad \left. < \frac{\sqrt{n}(\theta_0 - \theta)}{\sqrt{\theta(1-\theta)}} + z_{\alpha/2}\sqrt{\frac{\overline{Y}(1-\overline{Y})}{\theta(1-\theta)}}\right\} \\
&\approx 1 - Pr\left\{\frac{\sqrt{n}(\theta_0 - \theta)}{\sqrt{\theta(1-\theta)}} - z_{\alpha/2} < Z < \frac{\sqrt{n}(\theta_0 - \theta)}{\sqrt{\theta(1-\theta)}} + z_{\alpha/2}\right\} \\
&= 1 - \Phi\left(\frac{\sqrt{n}(\theta_0 - \theta)}{\sqrt{\theta(1-\theta)}} + z_{\alpha/2}\right) + \Phi\left(\frac{\sqrt{n}(\theta_0 - \theta)}{\sqrt{\theta(1-\theta)}} - z_{\alpha/2}\right),
\end{aligned}
$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal.

An R function to calculate approximate power

For the test of a single proportion

$$\text{Power} = 1 - \Phi\left(\frac{\sqrt{n}(\theta_0-\theta)}{\sqrt{\theta(1-\theta)}} + z_{\alpha/2}\right) + \Phi\left(\frac{\sqrt{n}(\theta_0-\theta)}{\sqrt{\theta(1-\theta)}} - z_{\alpha/2}\right)$$

```
Z2power = function(theta,n,theta0=0.50,alpha=0.05)
    {
    effect = sqrt(n)*(theta0-theta)/sqrt(theta*(1-theta))
    z = qnorm(1-alpha/2)
    Z2power = 1 - pnorm(effect+z) + pnorm(effect-z)
    Z2power
    } # End of function Z2power
```

Some numerical examples

`Z2power = function(theta,n,theta0=0.50,alpha=0.05)`

```
> Z2power(0.50,100) # Should be alpha = 0.05
[1] 0.05
>
> Z2power(0.55,100)
[1] 0.1713209
> Z2power(0.60,100)
[1] 0.5324209
> Z2power(0.65,100)
[1] 0.8819698
> Z2power(0.40,100)
[1] 0.5324209
> Z2power(0.55,500)
[1] 0.613098
> Z2power(0.55,1000)
[1] 0.8884346
```

Find smallest sample size needed to detect $\theta = 0.60$ as different from $\theta_0 = 0.50$ with probability at least 0.80

```
> samplesize = 1
> power=Z2power(theta=0.60,n=samplesize); power
[1] 0.05478667
> while(power < 0.80)
+ {
+ samplesize = samplesize+1
+ power = Z2power(theta=0.60,n=samplesize)
+ }
> samplesize
[1] 189
> power
[1] 0.8013024
```

What is required of the scientist

Who wants to select sample size by power analysis

The scientist must specify

- Parameter values that he or she wants to be able to detect as different from H_0 value.
- Desired power (probability of detection)

It's not always easy for a scientist to think in terms of the parameters of a statistical model.

Copyright Information