

Categorical Data Analysis¹

STA 312: Fall 2012

¹See last slide for copyright information.

Variables and Cases

- ▶ There are n **cases** (people, rats, factories, wolf packs) in a data set.
- ▶ A **variable** is a characteristic or piece of information that can be recorded for each case in the data set.
- ▶ For example cases could be patients in a hospital, and variables could be Age, Sex, Diagnosis, Have family doctor (Yes-No), Family history of heart disease (Yes-No), etc.

Variables can be Categorical, or Continuous

- ▶ Categorical: Gender, Diagnosis, Job category, Have family doctor, Family history of heart disease, 5-year survival (Y-N)
- ▶ Some categories are ordered (birth order, health status)
- ▶ Continuous: Height, Weight, Blood pressure
- ▶ Some questions:
 - ▶ Are all normally distributed variables continuous?
 - ▶ Are all continuous variables quantitative?
 - ▶ Are all quantitative variables continuous?
 - ▶ Are there really any data sets with continuous variables?

Variables can be Explanatory, or Response

- ▶ Explanatory variables are sometimes called “independent variables.”
- ▶ The x variables in regression are explanatory variables.
- ▶ Response variables are sometimes called “dependent variables.”
- ▶ The Y variable in regression is the response variable.
- ▶ Sometimes the distinction is not useful: Does each twin get cancer, Yes or No?

Our main interest is in categorical variables

- ▶ Especially categorical response variables
- ▶ In ordinary regression, outcomes are normally distributed, and so continuous.
- ▶ But often, outcomes of interest are categorical
 - ▶ Buy the product, or not
 - ▶ Marital status 5 years after graduation
 - ▶ Survive the operation, or not.
- ▶ Ordered categorical response variables, too: for example highest level of hockey ever played.

Distributions

We will mostly use

- ▶ Bernoulli
- ▶ Binomial
- ▶ Multinomial
- ▶ Poisson

The Poisson process

Why the Poisson distribution is such a useful model for count data

- ▶ Events happening randomly in space or time
- ▶ Independent increments
- ▶ For a small region or interval,
 - ▶ Chance of 2 or more events is negligible
 - ▶ Chance of an event roughly proportional to the size of the region or interval
- ▶ Then (solve a system of differential equations), the probability of observing x events in a region of size t is

$$\frac{e^{-\lambda t}(\lambda t)^x}{x!} \text{ for } x = 0, 1, \dots$$

Poisson process examples

Some variables that have a Poisson distribution

- ▶ Calls coming in to an emergency number
- ▶ Customers arriving in a given time period
- ▶ Number of raisins in a loaf of raisin bread
- ▶ Number of bomb craters in a region after a bombing raid, London WWII
- ▶ In a jar of peanut butter ...

Steps in the process of statistical analysis

One possible 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 parameters, perhaps 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 inference, draw conclusions about the example or problem

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

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

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

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

Find the MLE of π

Show your work

Maximize the log likelihood.

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

Setting the derivative to zero,

$$\begin{aligned}\frac{\sum_{i=1}^n y_i}{\pi} &= \frac{n - \sum_{i=1}^n y_i}{1 - \pi} \Rightarrow (1 - \pi) \sum_{i=1}^n y_i = \pi(n - \sum_{i=1}^n y_i) \\ &\Rightarrow \sum_{i=1}^n y_i - \pi \sum_{i=1}^n y_i = n\pi - \pi \sum_{i=1}^n y_i \\ &\Rightarrow \sum_{i=1}^n y_i = n\pi \\ &\Rightarrow \pi = \frac{\sum_{i=1}^n y_i}{n} = \bar{y} = p\end{aligned}$$

So it looks like the MLE is the sample proportion. Carrying out the second derivative test to be sure,

Second derivative test

$$\begin{aligned}\frac{\partial^2 \log \ell}{\partial \pi^2} &= \frac{\partial}{\partial \pi} \left(\frac{\sum_{i=1}^n y_i}{\pi} - \frac{n - \sum_{i=1}^n y_i}{1 - \pi} \right) \\ &= \frac{-\sum_{i=1}^n y_i}{\pi^2} - \frac{n - \sum_{i=1}^n y_i}{(1 - \pi)^2} \\ &= -n \left(\frac{1 - \bar{y}}{(1 - \pi)^2} + \frac{\bar{y}}{\pi^2} \right) < 0\end{aligned}$$

Concave down, maximum, and $\hat{\pi} = \bar{y} = p$.

Numerical estimate

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

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

Carry out a test to answer the question

Is there a difference in preference for the two blends?

Start by stating the null hypothesis

- ▶ $H_0 : \pi = 0.50$
- ▶ $H_1 : \pi \neq 0.50$
- ▶ A case could be made for a one-sided test, but we'll stick with two-sided.
- ▶ $\alpha = 0.05$ as usual.
- ▶ Central Limit Theorem says $\hat{\pi} = \bar{Y}$ is approximately normal with mean π and variance $\frac{\pi(1-\pi)}{n}$.

Several valid test statistics for $H_0 : \pi = \pi_0$ are available

Two of them are

$$Z_1 = \frac{\sqrt{n}(p - \pi_0)}{\sqrt{\pi_0(1 - \pi_0)}}$$

and

$$Z_2 = \frac{\sqrt{n}(p - \pi_0)}{\sqrt{p(1 - p)}}$$

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

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

Calculate the test statistic(s)

and the p -value(s)

```
> pi0 = .5; p = .6; n = 100
> Z1 = sqrt(n)*(p-pi0)/sqrt(pi0*(1-pi0)); Z1
[1] 2
> pval1 = 2 * (1-pnorm(Z1)); pval1
[1] 0.04550026
>
> Z2 = sqrt(n)*(p-pi0)/sqrt(p*(1-p)); 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? $\pi \neq 0.50$. *Specifically, $\pi > 0.50$.*
- ▶ 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 α 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 languages.

- ▶ “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

Approximately for large n ,

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

- ▶ Could express this as $p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$
- ▶ $z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$ 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.

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

In a report, you could say

- ▶ The estimated proportion preferring the new coffee bean blend is 0.60 ± 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 π .
- ▶ Does this mean $Pr\{0.504 < \pi < 0.696\} = 0.95$?
- ▶ No! The quantities 0.504, 0.696 and π are all constants, so $Pr\{0.504 < \pi < 0.696\}$ 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.
- ▶ So we sometimes say that we are “95% confident” that $0.504 < \pi < 0.696$.

Confidence intervals (regions) correspond to tests

Recall $Z_1 = \frac{\sqrt{n}(p-\pi_0)}{\sqrt{\pi_0(1-\pi_0)}}$ and $Z_2 = \frac{\sqrt{n}(p-\pi_0)}{\sqrt{p(1-p)}}$.

From the derivation of the confidence interval,

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

if and only if

$$p - z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} < \pi_0 < p + z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

- ▶ So the confidence interval consists of those parameter values π_0 for which $H_0 : \pi = \pi_0$ 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.
- ▶ There is a confidence interval corresponding to Z_1 too. Maybe it's better – See Chapter 1.
- ▶ In general, any test can be inverted to obtain a confidence region.

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

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.
- ▶ 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 π

$$Z_1 = \frac{\sqrt{n}(p - \pi_0)}{\sqrt{\pi_0(1 - \pi_0)}}$$

$$\begin{aligned} \text{Power} &= 1 - Pr\{-z_{\alpha/2} < Z_1 < z_{\alpha/2}\} \\ &= 1 - Pr\left\{-z_{\alpha/2} < \frac{\sqrt{n}(p - \pi_0)}{\sqrt{\pi_0(1 - \pi_0)}} < z_{\alpha/2}\right\} \\ &= \dots \\ &= 1 - Pr\left\{\frac{\sqrt{n}(\pi_0 - \pi)}{\sqrt{\pi(1 - \pi)}} - z_{\alpha/2}\sqrt{\frac{\pi_0(1 - \pi_0)}{\pi(1 - \pi)}} < \frac{\sqrt{n}(p - \pi)}{\sqrt{\pi(1 - \pi)}} \right. \\ &\quad \left. < \frac{\sqrt{n}(\pi_0 - \pi)}{\sqrt{\pi(1 - \pi)}} + z_{\alpha/2}\sqrt{\frac{\pi_0(1 - \pi_0)}{\pi(1 - \pi)}}\right\} \\ &\approx 1 - Pr\left\{\frac{\sqrt{n}(\pi_0 - \pi)}{\sqrt{\pi(1 - \pi)}} - z_{\alpha/2}\sqrt{\frac{\pi_0(1 - \pi_0)}{\pi(1 - \pi)}} < Z < \frac{\sqrt{n}(\pi_0 - \pi)}{\sqrt{\pi(1 - \pi)}} + z_{\alpha/2}\sqrt{\frac{\pi_0(1 - \pi_0)}{\pi(1 - \pi)}}\right\} \\ &= 1 - \Phi\left(\frac{\sqrt{n}(\pi_0 - \pi)}{\sqrt{\pi(1 - \pi)}} + z_{\alpha/2}\sqrt{\frac{\pi_0(1 - \pi_0)}{\pi(1 - \pi)}}\right) + \Phi\left(\frac{\sqrt{n}(\pi_0 - \pi)}{\sqrt{\pi(1 - \pi)}} - z_{\alpha/2}\sqrt{\frac{\pi_0(1 - \pi_0)}{\pi(1 - \pi)}}\right), \end{aligned}$$

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

$$\text{Power} = 1 - \Phi\left(\frac{\sqrt{n}(\pi_0 - \pi)}{\sqrt{\pi(1 - \pi)}} + z_{\alpha/2}\sqrt{\frac{\pi_0(1 - \pi_0)}{\pi(1 - \pi)}}\right) + \Phi\left(\frac{\sqrt{n}(\pi_0 - \pi)}{\sqrt{\pi(1 - \pi)}} - z_{\alpha/2}\sqrt{\frac{\pi_0(1 - \pi_0)}{\pi(1 - \pi)}}\right)$$

```
Z1power = function(pi,n,pi0=0.50,alpha=0.05)
{
  a = sqrt(n)*(pi0-pi)/sqrt(pi*(1-pi))
  b = qnorm(1-alpha/2) * sqrt(pi0*(1-pi0)/(pi*(1-pi)))
  Z1power = 1 - pnorm(a+b) + pnorm(a-b)
  Z1power
} # End of function Z1power
```

Some numerical examples

```
> Z1power(0.50,100)
[1] 0.05
>
> Z1power(0.55,100)
[1] 0.168788
> Z1power(0.60,100)
[1] 0.5163234
> Z1power(0.65,100)
[1] 0.8621995
> Z1power(0.40,100)
[1] 0.5163234
> Z1power(0.55,500)
[1] 0.6093123
> Z1power(0.55,1000)
[1] 0.8865478
```

Find smallest sample size needed to detect $\pi = 0.55$ as different from $\pi_0 = 0.50$ with probability at least 0.80

```
> samplesize = 50
> power=Z1power(pi=0.55,n=samplesize); power
[1] 0.1076602
> while(power < 0.80)
+ {
+ samplesize = samplesize+1
+ power = Z1power(pi=0.55,n=samplesize)
+ }
> samplesize; power
[1] 783
[1] 0.8002392
```

Find smallest sample size needed to detect $\pi = 0.60$ as different from $\pi_0 = 0.50$ with probability at least 0.80

```
> samplesize = 50
> power=Z1power(pi=0.60,n=samplesize); power
[1] 0.2890491
> while(power < 0.80)
+ {
+ samplesize = samplesize+1
+ power = Z1power(pi=0.60,n=samplesize)
+ }
> samplesize; power
[1] 194
[1] 0.8003138
```

Conclusions from the power analysis

- ▶ Detecting true $\pi = 0.60$ as different from 0.50 is a reasonable goal.
- ▶ Power with $n = 100$ is barely above one half – pathetic.
- ▶ As Fisher said, “To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination: he may be able to say what the experiment died of.”
- ▶ $n = 200$ is much better.
- ▶ How about $n = 250$?

```
> Z1power(pi=0.60,n=250)
```

```
[1] 0.8901088
```

It depends on what you can afford, but I like $n = 250$.

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 the scientist to think in terms of the parameters of a statistical model.

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/312f12>