Large Sample Chi-squared Tests: Part One[1]

STA442/2101 Fall 2013

---

[1]See last slide for copyright information.

Background Reading: Davison's *Statistical models*
It may be a little bit helpful.

- Section 2.2 on Convergence
- Pages 33-35 on the Delta method include the multivariate case
- The multinomial distribution is introduced as a homework problem, Problem 10 on p. 37.

Overview

Large-Sample Chi-square

Let $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then recall

$$(\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2(p)$$

It's true asymptotically too.

Using $(\mathbf{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2(p)$

Suppose

- $\sqrt{n}\,(\mathbf{T}_n - \boldsymbol{\theta}) \overset{d}{\to} \mathbf{T} \sim N\,(\mathbf{0}, \boldsymbol{\Sigma})$ and
- $\widehat{\boldsymbol{\Sigma}}_n \overset{p}{\to} \boldsymbol{\Sigma}$.

Then approximately as $n \to \infty$, $\mathbf{T}_n \sim N\left(\boldsymbol{\theta}, \frac{1}{n}\boldsymbol{\Sigma}\right)$, and

$$(\mathbf{T}_n - \boldsymbol{\theta})' \left(\frac{1}{n}\boldsymbol{\Sigma}\right)^{-1} (\mathbf{T}_n - \boldsymbol{\theta}) \sim \chi^2(p)$$

$$\|$$
$$n\,(\mathbf{T}_n - \boldsymbol{\theta})'\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{T}_n - \boldsymbol{\theta})$$
$$\approx\ n\,(\mathbf{T}_n - \boldsymbol{\theta})'\,\widehat{\boldsymbol{\Sigma}}_n^{-1}\,(\mathbf{T}_n - \boldsymbol{\theta})$$
$$\sim\ \chi^2(p)$$

Or we could be more precise

Suppose

- $\sqrt{n}\,(\mathbf{T}_n - \boldsymbol{\theta}) \overset{d}{\to} \mathbf{T} \sim N\,(\mathbf{0}, \boldsymbol{\Sigma})$ and
- $\widehat{\boldsymbol{\Sigma}}_n \overset{p}{\to} \boldsymbol{\Sigma}.$

Then $\widehat{\boldsymbol{\Sigma}}_n^{-1} \overset{p}{\to} \boldsymbol{\Sigma}^{-1}$, and by a Slutsky lemma,

$$\left(\begin{array}{c} \sqrt{n}\,(\mathbf{T}_n - \boldsymbol{\theta}) \\ \widehat{\boldsymbol{\Sigma}}_n^{-1} \end{array} \right) \overset{d}{\to} \left(\begin{array}{c} \mathbf{T} \\ \boldsymbol{\Sigma}^{-1} \end{array} \right).$$

By continuity,

$$\begin{aligned} & \left(\sqrt{n}(\mathbf{T}_n - \boldsymbol{\theta}) \right)' \widehat{\boldsymbol{\Sigma}}_n^{-1} \sqrt{n}(\mathbf{T}_n - \boldsymbol{\theta}) \\ = \ & n\,(\mathbf{T}_n - \boldsymbol{\theta})' \widehat{\boldsymbol{\Sigma}}_n^{-1} (\mathbf{T}_n - \boldsymbol{\theta}) \\ \overset{d}{\to} \ & \mathbf{T}' \boldsymbol{\Sigma}^{-1} \mathbf{T} \\ \sim \ & \chi^2(p) \end{aligned}$$

If $H_0 : \mathbf{L}\boldsymbol{\theta} = \mathbf{h}$ is true
Where $\mathbf{L}$ is $r \times p$ and of full row rank

Asymptotically, $\mathbf{LT}_n \sim N\left(\mathbf{L}\boldsymbol{\theta}, \frac{1}{n}\mathbf{L\Sigma L}'\right)$. So

$$(\mathbf{LT}_n - \mathbf{L}\boldsymbol{\theta})' \left(\frac{1}{n}\mathbf{L\Sigma L}'\right)^{-1} (\mathbf{LT}_n - \mathbf{L}\boldsymbol{\theta}) \sim \chi^2(r)$$

$$\|$$

$$n\left(\mathbf{LT}_n - \mathbf{h}\right)' \left(\mathbf{L\Sigma L}'\right)^{-1} \left(\mathbf{LT}_n - \mathbf{h}\right)$$

$$\approx \; n\left(\mathbf{LT}_n - \mathbf{h}\right)' \left(\mathbf{L}\widehat{\boldsymbol{\Sigma}}_n\mathbf{L}'\right)^{-1} \left(\mathbf{LT}_n - \mathbf{h}\right)$$

$$= W_n \; \sim \; \chi^2(r)$$

Or we could be more precise and use Slutsky lemmas.

Test of $H_0 : \mathbf{L}\boldsymbol{\theta} = \mathbf{h}$
Where $\mathbf{L}$ is $r \times p$ and of full row rank

$$W_n = n \left(\mathbf{L}\mathbf{T}_n - \mathbf{h}\right)' \left(\mathbf{L}\widehat{\boldsymbol{\Sigma}}_n\mathbf{L}'\right)^{-1} \left(\mathbf{L}\mathbf{T}_n - \mathbf{h}\right)$$

Distributed approximately as chi-squared with r degrees of freedom under H_0.

If $\mathbf{T}_n$ is the maximum likelihood estimator of $\boldsymbol{\theta}$, it's called a *Wald test* (and $\widehat{\boldsymbol{\Sigma}}_n$ has a special form).

Example: The `statclass` data

Fifty-eight students in a Statistics class took 8 quizzes, a midterm test and a final exam. They also had 9 computer assignments. The instructor wants to compare average performance on the four components of the grade.

- How about a model?
- Should we assume normality?
- Does it make sense to assume quiz marks independent of final exam marks?

Assume multivariate normality?

Histogram of computer
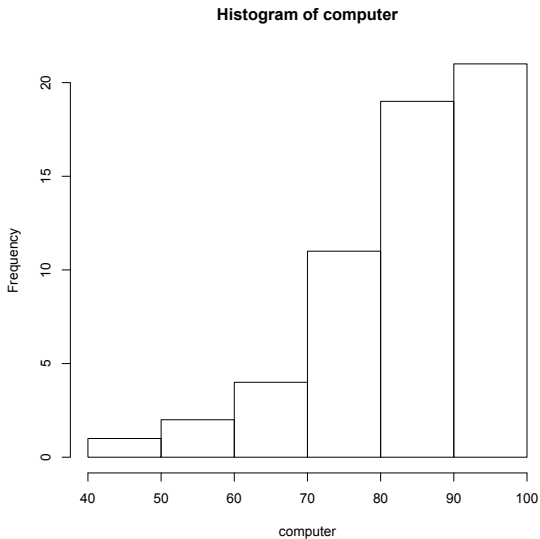
A model for the `statclass` data

Fifty-eight students in a Statistics class took 8 quizzes, a midterm test and a final exam. They also had 9 computer assignments.

Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be a random sample from an unknown distribution with mean $\mu = (\mu_1, \mu_2, \mu_3, \mu_4)'$ and covariance matrix $\mathbf{\Sigma}$.

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

Applying $W_n = n\left(\mathbf{L}\mathbf{T}_n - \mathbf{h}\right)'\left(\mathbf{L}\widehat{\boldsymbol{\Sigma}}_n\mathbf{L}'\right)^{-1}\left(\mathbf{L}\mathbf{T}_n - \mathbf{h}\right)$

To test $H_0 : \mathbf{L}\boldsymbol{\theta} = \mathbf{h}$

- Test is based on $\sqrt{n}\left(\mathbf{T}_n - \boldsymbol{\theta}\right) \overset{d}{\to} \mathbf{T} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$
- CLT says $\sqrt{n}\left(\overline{\mathbf{Y}}_n - \boldsymbol{\mu}\right) \overset{d}{\to} \mathbf{Y} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$
- So $\mathbf{T}_n = \overline{\mathbf{Y}}_n$ and $\boldsymbol{\theta} = \boldsymbol{\mu}$.
- Sample variance-covariance matrix is good enough for $\widehat{\boldsymbol{\Sigma}}_n \overset{p}{\to} \boldsymbol{\Sigma}$
- Write $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ as $\mathbf{L}\boldsymbol{\mu} = \mathbf{h}$

$H_0 : \mathbf{L}\boldsymbol{\theta} = \mathbf{h}$
To test equality of four means

$$
\begin{array}{cccc}
\mathbf{L} & \boldsymbol{\mu} & = & \mathbf{0} \\
\begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} & \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix} & = & \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}
\end{array}
$$

Read the data


```
> statclass = read.table("http://www.utstat.toronto.edu/~brunner/appliedf13/cod
> head(statclass)
  S E Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8 C1 C2 C3 C4 C5 C6 C7 C8 C9 MT  F
1 1 2  9  1  7  8  4  3  5  2  6 10 10 10  5  0  0  0  0 55 43
2 0 2 10 10  5  9 10  8  6  8 10 10  8  9  9  9  9 10 10 66 79
3 1 2 10 10  5 10 10 10  9  8 10 10 10 10 10 10 10  9 10 10 94 67
4 1 2 10 10  8  9 10  7 10  9 10 10 10  9 10 10  9 10 10 81 65
5 1 1 10  6  7  9  8  8  5  7 10  9 10  9  5  6  4  8 10 57 52
6 0 1 10  9  5  8  9  8  5  6  8  7  5  6 10  6  5  9  9 77 64
> attach(statclass)
The following object(s) are masked from 'package:base':

    F
```

Process the data a bit

```
> quiz = 10 * (Q1+Q2+Q3+Q4+Q5+Q6+Q7+Q8)/8
> computer = 10 * (C1+C2+C3+C4+C5+C6+C7+C8+C9)/9
> midterm = MT
> final = F
> datta = cbind(quiz,computer,midterm,final); head(round(datta))
     quiz computer midterm final
[1,]   49       46      55    43
[2,]   82       93      66    79
[3,]   90       99      94    67
[4,]   91       98      81    65
[5,]   75       79      57    52
[6,]   75       72      77    64
> ybar = apply(datta,2,mean); ybar
    quiz computer  midterm    final
72.58621 83.98467 68.87931 49.44828

> sigmahat = var(datta); sigmahat
              quiz   computer   midterm      final
quiz      120.66130  62.369765  60.31760  71.736993
computer   62.36977 134.894281  27.68233   6.272098
midterm    60.31760  27.682328 223.37114  99.633999
final      71.73699   6.272098  99.63400 272.777979
```

Calculate $W_n = n \left(\mathbf{LT}_n - \mathbf{h} \right)' \left(\mathbf{L} \widehat{\boldsymbol{\Sigma}}_n \mathbf{L}' \right)^{-1} \left(\mathbf{LT}_n - \mathbf{h} \right)$

```
> L = rbind(c(1,-1,0,0),
+           c(0,1,-1,0),
+           c(0,0,1,-1) )
> n = length(quiz); n
[1] 58
> Wn = n * t(L %*% ybar) %*% solve(L%*%sigmahat%*%t(L)) %*% L%*%ybar
> Wn
          [,1]
[1,] 176.8238
> Wn = as.numeric(Wn)
> pvalue = 1-pchisq(Wn,df=3); pvalue
[1] 0
```

Conclude that the four means are not all equal. Which ones are different from one another? Need follow-up tests.

The R function `Wtest`
"Estimated" asymptotic covariance matrix $\widehat{\mathbf{V}}_n = \frac{1}{n}\widehat{\mathbf{\Sigma}}_n$

$$
\begin{aligned}
W_n &= n\left(\mathbf{L}\mathbf{T}_n - \mathbf{h}\right)' \left(\mathbf{L}\widehat{\mathbf{\Sigma}}_n\mathbf{L}'\right)^{-1} \left(\mathbf{L}\mathbf{T}_n - \mathbf{h}\right) \\
&= \left(\mathbf{L}\mathbf{T}_n - \mathbf{h}\right)' \left(\mathbf{L}\widehat{\mathbf{V}}_n\mathbf{L}'\right)^{-1} \left(\mathbf{L}\mathbf{T}_n - \mathbf{h}\right)
\end{aligned}
$$

```
Wtest = function(L,Tn,Vn,h=0) # H0: L theta = h
# Note Vn is the estimated asymptotic covariance matrix of Tn,
# so it's Sigma-hat divided by n. For Wald tests based on numerical
# MLEs, Tn = theta-hat, and Vn is the inverse of the Hessian.
    {
    Wtest = numeric(3)
    names(Wtest) = c("W","df","p-value")
    r = dim(L)[1]
    W = t(L%*%Tn-h) %*% solve(L%*%Vn%*%t(L)) %*%
        (L%*%Tn-h)
    W = as.numeric(W)
    pval = 1-pchisq(W,r)
    Wtest[1] = W; Wtest[2] = r; Wtest[3] = pval
    Wtest
    } # End function Wtest
```

Illustrate the `Wtest` function

For $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$, got $W_n = 176.8238, df = 3, p \approx 0$.

```
> V = sigmahat / length(final)
> LL = rbind( c(1,-1,0,0),
+             c(0,1,-1,0),
+             c(0,0,1,-1) )
> Wtest(LL,ybar,V)
       W        df   p-value
176.8238    3.0000    0.0000
>
> ybar
    quiz computer  midterm     final
72.58621 83.98467 68.87931 49.44828
```

Is average quiz score different from midterm?

```
> L1 = rbind(c(1,0,-1,0)); n = length(final)
> Wtest(L=L1,Tn=ybar,Vn=sigmahat/n)
          W          df      p-value
3.56755878  1.00000000  0.05891887
```

Proceed to test all other pairwise differences between means.

The multinomial model

- A good source of examples
- Also directly useful in applications

Multinomial coefficient
For c categories

From n objects, number of ways to choose

- n_1 of type 1
- n_2 of type 2
    $\vdots$
- n_c of type c

$$\binom{n}{n_1 \; \cdots \; n_c} = \frac{n!}{n_1! \; \cdots \; n_c!}$$

Example of a multinomial coefficient
A counting problem

Of 30 graduating students who are seeking employment, how many ways are there for 15 to be employed in a job related to their field of study, 10 to be employed in a job unrelated to their field of study, and 5 unemployed?

$$\binom{30}{15\ 10\ 5} = 465,817,912,560$$

Multinomial Distribution
Denote by $M(n, \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_c)$

- Statistical experiment with c outcomes
- Repeated independently n times
- $Pr(\text{Outcome } j) = \theta_j$, $j = 1, \ldots, c$
- Number of times outcome j occurs is n_j, $j = 1, \ldots, c$
- An integer-valued *multivariate* distribution

$$P(n_1, \ldots, n_c) = \binom{n}{n_1 \; \cdots \; n_c} \theta_1^{n_1} \cdots \theta_c^{n_c},$$

where $0 \leq n_j \leq n$, $\sum_{j=1}^{c} n_j = n$, $0 < \theta_j < 1$, and $\sum_{j=1}^{c} \theta_j = 1$.

Example
Recent college graduates looking for a job

- Probability of job related to field of study = 0.60
- Probability of job unrelated to field of study = 0.30
- Probability of no job = 0.10

Of 30 randomly chosen students, what is probability that 15 are employed in a job related to their field of study, 10 are employed in a job unrelated to their field of study, and 5 are unemployed?

$$\binom{30}{15\ 10\ 5} 0.60^{15} 0.30^{10} 0.10^5 = \frac{4933527642332542053801}{381469726562500000000000} \approx 0.0129$$

Calculating multinomial probabilities with R

$$\binom{30}{15\ 10\ 5} 0.60^{15} 0.30^{10} 0.10^5 = \frac{4933527642332542053801}{381469726562500000000000} \approx 0.0129$$

```
> dmultinom(c(15,10,5), prob=c(.6, .3, .1))
[1] 0.01293295
```

There are actually $c-1$ variables and $c-1$ parameters
In the multinomial with c categories

$$
\begin{aligned}
P(n_1,\ldots,n_{c-1}) &= \frac{n!}{n_1! \, \cdots \, n_{c-1}!(n-\sum_{j=1}^{c-1} n_j)!} \\
&\quad \times \theta_1^{n_1}\cdots\theta_{c-1}^{n_{c-1}}(1-\sum_{j=1}^{c-1}\theta_j)^{n-\sum_{j=1}^{c-1} n_j}
\end{aligned}
$$

Useful fact
Proof omitted

If you combine the categories of a multinomial (adding the frequencies), the result is still multinomial (adding the probabilities).

- Roll a die 100, times, observe the events 1, 3, 5, even.
- Collapse into just 2 categories, and the result is binomial.

Hypothetical data file
Let $Y_{i,j}$ be indicators for category membership, $i = 1, \ldots, n$ and $j = 1, \ldots, c$

Case	Job	Y_1	Y_2	Y_3
1	1	1	0	0
2	3	0	0	1
3	2	0	1	0
4	1	1	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	2	0	1	0
Total		$\sum_{i=1}^n y_{i,1}$	$\sum_{i=1}^n y_{i,2}$	$\sum_{i=1}^n y_{i,3}$

Note that

- A real data file will almost never have the redundant variables Y_1, Y_2 and Y_3.
- $\sum_{i=1}^n y_{i,j} = n_j$

Lessons from the data file

- Cases (n of them) are independent $M(1, \boldsymbol{\theta})$, so $E(Y_{i,j}) = \theta_j$.
- Column totals n_j count the number of times each category occurs: Joint distribution is $M(n, \boldsymbol{\theta})$
- If you make a frequency table (frequency distribution)
    - The n_j counts are the cell frequencies!
    - They are random variables, and now we know their joint distribution.
    - Each individual (marginal) table frequency is $B(n, \theta_j)$.
    - Expected value of cell frequency j is $E(n_j) = n\theta_j$
- Tables of 2 and or more dimensions present no problems; form combination variables.

Example of a frequency table
For the Jobs data

Job Category	Frequency	Percent
Employed in field	106	53
Employed outside field	74	37
Unemployed	20	10
Total	200	100.0

Estimation based on $E(Y_{i,j}) = \theta_j$

$$
E(\overline{\mathbf{Y}}) = E \begin{pmatrix} \frac{1}{n} \sum_{i=1}^{n} Y_{i,1} \\[6pt] \frac{1}{n} \sum_{i=1}^{n} Y_{i,2} \\ \vdots \\ \frac{1}{n} \sum_{i=1}^{n} Y_{i,c} \end{pmatrix} = E \begin{pmatrix} \overline{Y}_1 \\[6pt] \overline{Y}_2 \\ \vdots \\ \overline{Y}_c \end{pmatrix} = \begin{pmatrix} \theta_1 \\[6pt] \theta_2 \\ \vdots \\ \theta_c \end{pmatrix}
$$

The sample mean vector is

- A vector of sample proportions.
- Unbiased for $\boldsymbol{\theta}$.
- Consistent for $\boldsymbol{\theta}$ because by the Law of Large Numbers,

$$
\overline{\mathbf{Y}}_n \overset{a.s.}{\rightarrow} \boldsymbol{\theta}
$$

Multivariate Central Limit Theorem

To avoid a singular covariance matrix, let

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_{c-1} \end{pmatrix} \quad \text{and} \quad \overline{\mathbf{Y}}_n = \begin{pmatrix} \overline{Y}_1 \\ \vdots \\ \overline{Y}_{c-1} \end{pmatrix}$$

Multivariate Central Limit Theorem says let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be a random sample from a distribution with expected value vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Then

$$\sqrt{n}(\overline{\mathbf{Y}}_n - \boldsymbol{\mu}) \xrightarrow{d} \mathbf{Y} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}\right).$$

Asymptotically speaking
That is, approximately for large n

$\overline{\mathbf{Y}}_n$ is normal with

- Mean $\boldsymbol{\mu}$ and
- Covariance matrix $\frac{1}{n}\boldsymbol{\Sigma}$.

$$\sqrt{n}(\overline{\mathbf{Y}}_n - \boldsymbol{\mu}) \xrightarrow{d} \mathbf{Y} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$$

For multinomial data,

- Have $\boldsymbol{\mu} = \boldsymbol{\theta}$
- Because $\mathbf{Y}_i \sim M(1, \boldsymbol{\theta})$, the diagonal elements of $\boldsymbol{\Sigma}$ are $Var(Y_{i,j}) = \theta_j(1 - \theta_j)$.
- What about the off-diagonal elements?

$$
\begin{aligned}
Cov(Y_{i,j}, Y_{i,k}) &= E(Y_{i,j}Y_{i,k}) - E(Y_{i,j})E(Y_{i,k}) \\
&= \quad 0 \quad - \quad \theta_j \quad \theta_k
\end{aligned}
$$

Covariance matrix for $c = 4$ categories

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \begin{pmatrix} \theta_1(1 - \theta_1) & -\theta_1\theta_2 & -\theta_1\theta_3 \\ -\theta_1\theta_2 & \theta_2(1 - \theta_2) & -\theta_2\theta_3 \\ -\theta_1\theta_3 & -\theta_2\theta_3 & \theta_3(1 - \theta_3) \end{pmatrix}$$

- Last category is dropped because it's redundant.
- Consistent estimator of $\boldsymbol{\Sigma}$ is easy.
- All you need is the frequency table.

A testing problem

Administrators at a vocational college recognize that the percentage of students who are unemployed after graduation will vary depending upon economic conditions, but they claim that still, about twice as many students will be employed in a job related to their field of study, compared to those who get an unrelated job. To test this hypothesis, they select a random sample of 200 students from the most recent class, and observe 106 employed in a job related to their field of study, 74 employed in a job unrelated to their field of study, and 20 unemployed. Test the hypothesis using a large-sample chi-squared test, being guided by the usual 0.05 significance level. State your conclusions in symbols and words.

Some questions to help guide us through the problem

- What is the model?

$$\mathbf{Y}_1, \ldots, \mathbf{Y}_n \overset{i.i.d.}{\sim} M(1, (\theta_1, \theta_2, \theta_3))$$

- What is the null hypothesis, in symbols?

$$H_0 : \theta_1 = 2\theta_2$$

- Give the null hypothesis in matrix form as $H_0 : \mathbf{L}\boldsymbol{\theta} = \mathbf{h}$

$$(1, -2) \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = (0)$$

Last category is dropped because it's redundant.

Test statistic
For the Wald-like test

Suppose

- $\sqrt{n}\left(\mathbf{T}_n - \boldsymbol{\theta}\right) \overset{d}{\to} \mathbf{T} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$ and
- $\widehat{\boldsymbol{\Sigma}}_n \overset{p}{\to} \boldsymbol{\Sigma}$.
- $H_0 : \mathbf{L}\boldsymbol{\theta} = \mathbf{h}$, where $\mathbf{L}$ is $r \times p$ of full row rank.

Then

$$W_n = n\left(\mathbf{L}\mathbf{T}_n - \mathbf{h}\right)'\left(\mathbf{L}\widehat{\boldsymbol{\Sigma}}_n\mathbf{L}'\right)^{-1}\left(\mathbf{L}\mathbf{T}_n - \mathbf{h}\right) \overset{d}{\to} W \sim \chi^2(r).$$

Calculating the test statistic with R
Frequencies were in=106, out=74, un=20

```
> # Easy to modify for any multinomial problem. Omit one category.
> freq = c(106,74,20); n = sum(freq)
> ybar = (freq/n)[1:2] # Just the first 2
> p = length(ybar)
> Sighat = matrix(nrow=p,ncol=p) # Empty matrix
> for(i in 1:p) { for (j in 1:p) Sighat[i,j] = -ybar[i]*ybar[j] }
> for(i in 1:p) Sighat[i,i] = ybar[i]*(1-ybar[i])
> Sighat
        [,1]     [,2]
[1,]  0.2491 -0.1961
[2,] -0.1961  0.2331
>
> LL = rbind(c(1,-2))
> source("http://www.utstat.utoronto.ca/~brunner/appliedf13/code_n_data
/lecture/Wtest.txt")
> Wtest(LL,ybar,Sighat/n)
          W           df      p-value
4.48649474 1.00000000 0.03416366
```

Conclude more than predicted are employed outside their field of study.

Beer study

Under carefully controlled conditions, 120 beer drinkers each tasted 6 beers and indicated which one they liked best. Here are the numbers preferring each beer.

	Preferred Beer					
	1	2	3	4	5	6
Frequency	30	24	22	28	9	7

- State a reasonable model.

$$\mathbf{Y}_1, \ldots, \mathbf{Y}_n \overset{i.i.d.}{\sim} M(1, (\theta_1, \theta_2, \ldots, \theta_6))$$

The first question

Are preferences for the 6 beers is different in the population from which this sample was taken?

- State the null hypothesis.

$$\theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta_5 = \frac{1}{6}$$

- Give the null hypothesis in matrix form as $H_0 : \mathbf{L}\boldsymbol{\theta} = \mathbf{h}$

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \end{pmatrix} = \begin{pmatrix} 1/6 \\ 1/6 \\ 1/6 \\ 1/6 \\ 1/6 \end{pmatrix}$$

- What are the degrees of freedom of the test?

$$df = 5$$

Test for equal probabilities

```
> # Estimated covariance matrix for a multinomial:
> # Omit one category.
> freq = c(30,24,22,28,9,7); n = sum(freq)
> ybar = (freq/n)[1:5] # Omit lat category
> p = length(ybar)
> Sighat = matrix(nrow=p,ncol=p) # Empty matrix
> for(i in 1:p) { for (j in 1:p) Sighat[i,j] = -ybar[i]*ybar[j] }
> for(i in 1:p) Sighat[i,i] = ybar[i]*(1-ybar[i])
>
> # L matrix is the 5x5 identity for this hypothesis
> alleq = numeric(5) + 1/6; alleq
[1] 0.1666667 0.1666667 0.1666667 0.1666667 0.1666667
> test1 = Wtest(L=diag(5),Tn=ybar,Vn=Sighat/n,h=alleq); test1
            W              df         p-value
4.405483e+01 5.000000e+00 2.257610e-08
>
```

Compare Pearson chi-squared test for equal probabilities

$X^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$

```
> round(test1,3)
      W       df p-value
 44.055    5.000   0.000

> expected = numeric(6) + n/6
> rbind(freq,expected)
         [,1] [,2] [,3] [,4] [,5] [,6]
freq       30   24   22   28    9    7
expected   20   20   20   20   20   20
> X2 = sum((freq-expected)^2/expected)
> X2
[1] 23.7
>
> 1-pchisq(X2,5)
[1] 0.0002479085
```

Take a look
Where is the lack of fit coming from?

	Preferred Beer					
	1	2	3	4	5	6
Frequency	30	24	22	28	9	7
Expected	20	20	20	20	20	20
Residual	10	4	2	8	-11	-13
St. Residual	2.11	0.91	0.47	1.73	-3.81	-5.06

To standardize the a residual, divide it by the estimated
standard deviation.

A follow-up question

It seems that the first 4 beers are lagers and the last two are ales. The client says no one would expect preference for lagers and ales to be the same. So let's test whether preference for the 4 lagers is different, and at the same time, whether preference for the 2 ales is different.

- Why not just look at the standardized residuals, which are Z-tests?

- Give the null hypothesis in symbols.

$$\theta_1 = \theta_2 = \theta_3 = \theta_4$$

$$\theta_5 = 1 - \theta_1 - \theta_2 - \theta_3 - \theta_4 - \theta_5$$

Null hypothesis in matrix form
For $\theta_1 = \theta_2 = \theta_3 = \theta_4$ and $\theta_1 + \theta_2 + \theta_3 + \theta_4 + 2\theta_5 = 1$

$$\begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 1 & 1 & 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

There are infinitely many right answers.

Test for differences among ales and differences between lagers, simultaneously

```
> # Test for diff among ales and diff between lagers
> L2 = rbind( c(1,-1, 0, 0, 0),
+             c(0, 1,-1, 0, 0),
+             c(0, 0, 1,-1, 0),
+             c(1, 1, 1, 1, 2)  )
> h2 = cbind(c(0,0,0,1))
> Wtest(L2,ybar,Sighat/n,h2)
        W         df   p-value
1.8240899 4.0000000 0.7680721
```

Test mean preference for ales vs. mean preference for lagers, just for completeness

The null hypothesis is

$$\frac{1}{4}\left(\theta_1 + \theta_2 + \theta_3 + \theta_4\right) = \frac{1}{2}\left(\theta_5 + 1 - \sum_{j=1}^{5}\theta_j\right)$$

$$\Leftrightarrow \quad \theta_1 + \theta_2 + \theta_3 + \theta_4 = \frac{2}{3}$$

```
> # Average ale vs. average lager
> L3 = rbind(c(1,1,1,1,0)); h3=2/3
> Wtest(L3,ybar,Sighat/n,h3)
           W            df     p-value
4.153846e+01 1.000000e+00 1.155747e-10
```

State the conclusion in plain language

Consumers tend to prefer ales over lagers. Differences between ales and differences between lagers are small enough to be attributed to chance.

Another application: Mean index numbers

In a study of consumers' opinions of 5 popular TV programmes, 240 consumers who watch all the shows at least once a month completed a computerized interview. On one of the screens, they indicated how much they enjoyed each programme by mouse-clicking on a 10cm line. One end of the line was labelled "Like very much," and the other end was labelled "Dislike very much." So each respondent contributed 5 ratings, on a continuous scale from zero to ten.

The study was commissioned by the producers of one of the shows, which will be called "Programme E." Ratings of Programmes A through D were expressed as percentages of the rating for Programme E, and these were described as "Liking indexed to programme E."

In statistical language

We have $X_{i,1}, \ldots X_{i,5}$ for $i = 1, \ldots, n$, and we calculate

$$Y_{i,j} = 100 \frac{X_{i,j}}{X_{i,5}}$$

- We want confidence intervals for the 4 mean index numbers, and tests of differences between means.
- Observations from the same respondent are definitely not independent.
- What is the distribution?
- What is a reasonable model?

Model

Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be a random sample from an unknown multivariate distribution F with expected value $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

One way to think about it is

- The parameter is the unknown distribution F.
- The parameter space is a space of distribution functions.
- $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are *functions* of F.
- We're only interested in $\boldsymbol{\mu}$.

We have the tools we need

- $\sqrt{n}(\overline{\mathbf{Y}}_n - \boldsymbol{\mu}) \xrightarrow{d} \mathbf{Y} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ and
- For $\widehat{\boldsymbol{\Sigma}}_n \xrightarrow{p} \boldsymbol{\Sigma}$, use the sample covariance matrix.
- $H_0 : \mathbf{L}\boldsymbol{\mu} = \mathbf{h}$

$$W_n = n\left(\mathbf{L}\overline{\mathbf{Y}}_n - \mathbf{h}\right)'\left(\mathbf{L}\widehat{\boldsymbol{\Sigma}}_n\mathbf{L}'\right)^{-1}\left(\mathbf{L}\overline{\mathbf{Y}}_n - \mathbf{h}\right)$$

Read the data

```
> Y = read.table("http://www.utstat.toronto.edu/~brunner/appliedf13/code_n_data/lecture/TVshows.data")

> Y[1:4,]
      A     B     C     D
1 101.3  81.0 101.8  89.6
2  94.0  85.3  76.3 100.8
3 145.4 138.7 151.0 148.3
4  72.0  86.1  96.1  96.3
> n = dim(Y)[1]; n
[1] 240
```

Confidence intervals

```
> ave = apply(Y,2,mean); ave
        A         B         C         D
101.65958  98.50167  99.39958 103.94167

> v = apply(Y,2,var) # Sample variances with n-1
> stderr = sqrt(v/n)
> me95 = 1.96*stderr
> lower95 = ave-me95
> upper95 = ave+me95
> Z = (ave-100)/stderr
> rbind(ave,marginerror95,lower95,upper95,Z)
                        A          B           C          D
ave            101.659583  98.501667  99.3995833 103.941667
marginerror95    1.585652   1.876299   1.7463047   1.469928
lower95        100.073931  96.625368  97.6532786 102.471739
upper95        103.245236 100.377966 101.1458880 105.411594
Z                2.051385  -1.565173  -0.6738897   5.255814
```

What if we "assume" normality?

```
> rbind(ave,lower95,upper95,Z)
                 A          B          C          D
ave      101.659583  98.501667  99.3995833 103.941667
lower95  100.073931  96.625368  97.6532786 102.471739
upper95  103.245236 100.377966 101.1458880 105.411594
Z          2.051385  -1.565173  -0.6738897   5.255814
> attach(Y) # So A, B, C, D are available
> t.test(A,mu=100)

One Sample t-test

data:  A
t = 2.0514, df = 239, p-value = 0.04132
alternative hypothesis: true mean is not equal to 100
95 percent confidence interval:
 100.0659 103.2533
sample estimates:
mean of x
 101.6596
```

Test equality of means

```
> S = var(Y); S
         A         B         C        D
A 157.0779 110.77831 106.56220 109.6234
B 110.7783 219.93950  95.66686 100.3585
C 106.5622  95.66686 190.51937 106.2501
D 109.6234 100.35851 106.25006 134.9867
> cor(Y)
          A         B         C        D
A 1.0000000 0.5959991 0.6159934 0.7528355
B 0.5959991 1.0000000 0.4673480 0.5824479
C 0.6159934 0.4673480 1.0000000 0.6625431
D 0.7528355 0.5824479 0.6625431 1.0000000
>
> L4 = rbind( c(1,-1, 0, 0),
+             c(0, 1,-1, 0),
+             c(0, 0, 1,-1)  )
> Wtest(L=L4,Tn=ave,Vn=S/n)
           W          df     p-value
7.648689e+01 3.000000e+00 2.220446e-16
```

Between vs. within cases

We have been comparing means. The means represent the average of what happens under different conditions.

- Within cases: A case (individual, sampling unit) contributes data in all the conditions.
- Between cases: Different cases in different conditions.
- Think of a one-sample vs. 2-sample t-test
- All the examples so far have been within cases.

Independent groups (Between cases)
Like a one-factor ANOVA

- Have n cases, separated into p groups: Maybe occupation of main wage earner in family.

- $n_1 + n_2 + \cdots + n_p = n$

- Response variable is either binary or quantity of something, like annual energy consumption.

- No reason to believe normality.

- No reason to believe equal variances.

- $H_0 : \mathbf{L}\boldsymbol{\mu} = \mathbf{h}$

- For example, $H_0 : \mu_1 = \ldots = \mu_p$

- Or $\mu_2 = \mu_7$

Basic Idea

The p sample means are independent random variables.
Asymptotically,

- $\overline{Y}_j \sim N(\mu_j, \frac{\sigma_j^2}{n_j})$

- The $p \times 1$ random vector $\overline{\mathbf{Y}}_n \sim N(\boldsymbol{\mu}, \mathbf{V}_n)$,

- Where $\mathbf{V}_n$ is a $p \times p$ diagonal matrix with jth diagonal element $\frac{\sigma_j^2}{n_j}$.

- $\mathbf{L}\overline{\mathbf{Y}}_n \sim N_r(\mathbf{L}\boldsymbol{\mu}, \mathbf{L}\mathbf{V}_n\mathbf{L}')$

- "Estimate" $\mathbf{V}_n$ with the diagonal matrix $\widehat{\mathbf{V}}_n$, jth diagonal element $\frac{\widehat{\sigma}_j^2}{n_j}$

- And if $H_0 : \mathbf{L}\boldsymbol{\mu} = \mathbf{h}$ is true,

$$
W_n = \left(\mathbf{L}\overline{\mathbf{Y}}_n - \mathbf{h}\right)' \left(\mathbf{L}\widehat{\mathbf{V}}_n\mathbf{L}'\right)^{-1} \left(\mathbf{L}\overline{\mathbf{Y}}_n - \mathbf{h}\right) \sim \chi^2(r)
$$

One little technical issue

- More than one n_j is going to infinity.
- The rates at which they go to infinity can't be too different.
- In particular, if $n = n_1 + n_2 + \cdots + n_p$,
- Then each $\frac{n_j}{n}$ must converge to a non-zero constant (in probability).

Compare High School marks for students at 3 campuses

Campus	n	Mean	Standard Deviation
SG	3906	84.94	5.59
UTM	1583	79.68	5.82
UTSC	1849	79.96	5.98

Compute $W_n = \left(\mathbf{L\overline{Y}}_n - \mathbf{h}\right)' \left(\mathbf{L\widehat{V}}_n\mathbf{L}'\right)^{-1} \left(\mathbf{L\overline{Y}}_n - \mathbf{h}\right)$

$H_0 : \mu_1 = \mu_2 = \mu_3$

Campus	n	Mean	Standard Deviation
SG	3906	84.94	5.59
UTM	1583	79.68	5.82
UTSC	1849	79.96	5.98

```
> source("http://www.utstat.utoronto.ca/~brunner/appliedf13/code_n_data/lecture/Wtest.txt")

> n = c(3906,1583,1849)
> ybar = c(84.94,79.68,79.96)
> Vhat = diag(c(5.59,5.82,5.98)^2/n); Vhat
            [,1]       [,2]       [,3]
[1,] 0.008000026 0.0000000 0.0000000
[2,] 0.000000000 0.0213976 0.0000000
[3,] 0.000000000 0.0000000 0.0193404
> L1 = rbind(c(1,-1,0),
+            c(0,1,-1)  )
> Wtest(L1,ybar,Vhat)
      W       df p-value
1441.58    2.00    0.00
```

Test difference between UTM and UTSC

Campus	n	Mean	Standard Deviation
SG	3906	84.94	5.59
UTM	1583	79.68	5.82
UTSC	1849	79.96	5.98

```
> # UTM vs. UTSC
> Wtest(rbind(c(0,1,-1)),ybar,Vhat)
        W        df   p-value
1.9244931 1.0000000 0.1653622
```

The Delta Method

If a sequence of random vectors converges to something nice in distribution, the Delta Method helps with the convergence of *functions* of those random vectors.

The Jacobian

- Univariate version of the Delta method says
  $\sqrt{n}\left(g(T_n) - g(\theta)\right) \xrightarrow{d} g'(\theta)\, T$
- In the multivariate version, the derivative is replaced by a matrix of partial derivatives, with i, j element $\frac{\partial g_i}{\partial x_j}$.
- Say the function g maps $\mathbb{R}^3$ into $\mathbb{R}^2$. Then

$$\dot{g}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial}{\partial \theta_1} g_1(\boldsymbol{\theta}) & \frac{\partial}{\partial \theta_2} g_1(\boldsymbol{\theta}) & \frac{\partial}{\partial \theta_3} g_1(\boldsymbol{\theta}) \\[2mm] \frac{\partial}{\partial \theta_1} g_2(\boldsymbol{\theta}) & \frac{\partial}{\partial \theta_2} g_2(\boldsymbol{\theta}) & \frac{\partial}{\partial \theta_3} g_2(\boldsymbol{\theta}) \end{pmatrix}$$

  where $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)'$.
- It's a Jacobian.

The Multivariate Delta Method

Let $g : \mathbb{R}^d \to \mathbb{R}^k$ be such that the elements of $\dot{g}(\mathbf{x}) = \left[\frac{\partial g_i}{\partial x_j} \right]_{k \times d}$ are continuous in a neighborhood of $\boldsymbol{\theta} \in \mathbb{R}^d$.

- If $\mathbf{T}_n$ is a sequence of d-dimensional random vectors such that $\sqrt{n}(\mathbf{T}_n - \boldsymbol{\theta}) \xrightarrow{d} \mathbf{T}$, then $\sqrt{n}(g(\mathbf{T}_n) - g(\boldsymbol{\theta})) \xrightarrow{d} \dot{g}(\boldsymbol{\theta})\mathbf{T}$.

- In particular, if $\sqrt{n}(\mathbf{T}_n - \boldsymbol{\theta}) \xrightarrow{d} \mathbf{T} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, then $\sqrt{n}(g(\mathbf{T}_n) - g(\boldsymbol{\theta})) \xrightarrow{d} \mathbf{Y} \sim N(\mathbf{0}, \dot{g}(\boldsymbol{\theta})\boldsymbol{\Sigma}\dot{g}(\boldsymbol{\theta})')$.

Example: The Japanese car study

In a study commissioned by the Honda motor company[2], a
random sample of 1200 Canadians who purchased a new
Japanese car within the past 12 months gave the make of the
most recent one they bought. Company executives like to look
at percent differences, compared to choice of their own brand.

```
> car = c(316, 414, 138, 254,  28,  50)
> names(car) = c("Honda", "Toyota", "Nissan", "Mazda", "Mitsubishi", "Suburu")
> perdiff = round(100 * (car-car[1])/car[1],1)

> car
    Honda     Toyota     Nissan     Mazda Mitsubishi     Suburu
      316        414        138        254         28         50
> perdiff
    Honda     Toyota     Nissan     Mazda Mitsubishi     Suburu
      0.0       31.0      -56.3      -19.6      -91.1      -84.2
```

---

[2]Not really. All the numbers are made up, based loosely on figures I
found on the Internet.

Multinomial model with 6 categories

- Make Honda the "other" category, for symmetry.
- Parameters are $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_5)'$.
- For compactness, let $\theta_0 = 1 - \sum_{k=1}^{5} \theta_k$, the probability of buying a Honda.
- Estimate $\boldsymbol{\theta}$ with the vector of sample proportions $\overline{\mathbf{Y}} = (\overline{Y}_1, \ldots, \overline{Y}_5)'$.
- The *functions* of $\boldsymbol{\theta}$ that interest us are the population percent differences:

$$g_i(\boldsymbol{\theta}) = 100 \left(\frac{\theta_i - \theta_0}{\theta_0} \right) = 100 \left(\frac{\theta_i}{\theta_0} - 1 \right)$$

for $i = 1, \ldots, 5$.

The function $g(\boldsymbol{\theta})$
Written a better way

$$g(\boldsymbol{\theta}) = \begin{pmatrix} g_1(\boldsymbol{\theta}) \\ g_2(\boldsymbol{\theta}) \\ g_3(\boldsymbol{\theta}) \\ g_4(\boldsymbol{\theta}) \\ g_5(\boldsymbol{\theta}) \end{pmatrix} = \begin{pmatrix} 100\left(\frac{\theta_1}{\theta_0} - 1\right) \\ 100\left(\frac{\theta_2}{\theta_0} - 1\right) \\ 100\left(\frac{\theta_3}{\theta_0} - 1\right) \\ 100\left(\frac{\theta_4}{\theta_0} - 1\right) \\ 100\left(\frac{\theta_5}{\theta_0} - 1\right) \end{pmatrix}$$

The Jacobian $\dot{\mathrm{g}}(\boldsymbol{\theta})$ is a 5×5 matrix of partial derivatives.

That's only because $g : \mathbb{R}^5 \to \mathbb{R}^5$. Usually $\dot{\mathrm{g}}(\boldsymbol{\theta})$ has more columns than rows.

Using $\theta_0 = 1 - \sum_{k=1}^{5} \theta_k$ and noting $\frac{\partial \theta_0}{\partial \theta_j} = -1$,

$$
\frac{\partial g_i}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} 100 \left(\frac{\theta_i}{\theta_0} - 1 \right) = \begin{cases} 100 \frac{\theta_i + \theta_0}{\theta_0^2} & \text{if } i = j \\[2ex] 100 \frac{\theta_i}{\theta_0^2} & \text{if } i \neq j \end{cases}
$$

The Jacobian

$$\dot{g}(\boldsymbol{\theta}) = \frac{100}{\theta_0^2} \begin{pmatrix} \theta_1 + \theta_0 & \theta_1 & \theta_1 & \theta_1 & \theta_1 \\ \theta_2 & \theta_2 + \theta_0 & \theta_2 & \theta_2 & \theta_2 \\ \theta_3 & \theta_3 & \theta_3 + \theta_0 & \theta_3 & \theta_3 \\ \theta_4 & \theta_4 & \theta_4 & \theta_4 + \theta_0 & \theta_4 \\ \theta_5 & \theta_5 & \theta_5 & \theta_5 & \theta_5 + \theta_0 \end{pmatrix}$$

The asymptotic covariance matrix

The Delta Method says

$$\sqrt{n}(g(\overline{\mathbf{Y}}_n) - g(\boldsymbol{\theta})) \xrightarrow{d} \mathbf{Y} \sim N(\mathbf{0}, \dot{g}(\boldsymbol{\theta})\boldsymbol{\Sigma}(\boldsymbol{\theta})\dot{g}(\boldsymbol{\theta})')$$

So the asymptotic covariance matrix of $g(\overline{\mathbf{Y}}_n)$ (the percent differences from Honda) is $\frac{1}{n}\dot{g}(\boldsymbol{\theta})\boldsymbol{\Sigma}(\boldsymbol{\theta})\dot{g}(\boldsymbol{\theta})'$, where

$$\dot{g}(\boldsymbol{\theta}) = \frac{100}{\theta_0^2} \begin{pmatrix} \theta_1 + \theta_0 & \theta_1 & \theta_1 & \theta_1 & \theta_1 \\ \theta_2 & \theta_2 + \theta_0 & \theta_2 & \theta_2 & \theta_2 \\ \theta_3 & \theta_3 & \theta_3 + \theta_0 & \theta_3 & \theta_3 \\ \theta_4 & \theta_4 & \theta_4 & \theta_4 + \theta_0 & \theta_4 \\ \theta_5 & \theta_5 & \theta_5 & \theta_5 & \theta_5 + \theta_0 \end{pmatrix}$$

and

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \begin{pmatrix} \theta_1(1 - \theta_1) & -\theta_1\theta_2 & -\theta_1\theta_3 & -\theta_1\theta_4 & -\theta_1\theta_5 \\ -\theta_1\theta_2 & \theta_2(1 - \theta_2) & -\theta_2\theta_3 & -\theta_2\theta_4 & -\theta_2\theta_5 \\ -\theta_1\theta_3 & -\theta_2\theta_3 & \theta_3(1 - \theta_3) & -\theta_3\theta_4 & -\theta_3\theta_5 \\ -\theta_1\theta_4 & -\theta_2\theta_4 & -\theta_3\theta_4 & \theta_4(1 - \theta_4) & -\theta_4\theta_5 \\ -\theta_1\theta_5 & -\theta_2\theta_5 & -\theta_3\theta_5 & -\theta_4\theta_5 & \theta_5(1 - \theta_5) \end{pmatrix}$$

Estimate θ_j with $\overline{Y}_j$.

Calculate $\mathbf{\Sigma}(\widehat{\boldsymbol{\theta}})$

```
> # Calculate estimated Sigma(theta). Did this kind of thing earlier.
> p = length(ybar)
> Sighat = matrix(nrow=p,ncol=p) # Empty matrix
> for(i in 1:p) { for (j in 1:p) Sighat[i,j] = -ybar[i]*ybar[j] }
> for(i in 1:p) Sighat[i,i] = ybar[i]*(1-ybar[i])
> Sighat
           [,1]          [,2]          [,3]          [,4]          [,5]
[1,]  0.225975 -0.039675000 -0.073025000 -0.0080500000 -0.0143750000
[2,] -0.039675  0.101775000 -0.024341667 -0.0026833333 -0.0047916667
[3,] -0.073025 -0.024341667  0.166863889 -0.0049388889 -0.0088194444
[4,] -0.008050 -0.002683333 -0.004938889  0.0227888889 -0.0009722222
[5,] -0.014375 -0.004791667 -0.008819444 -0.0009722222  0.0399305556
```

Calculate $\dot{g}(\widehat{\boldsymbol{\theta}})$

```
> # Calculate estimated gdot
> ybar0 = car[1]/n # Proportion buying Honda
> gdot = matrix(nrow=p,ncol=p) # Empty matrix
> for(i in 1:p) gdot[i,] = numeric(p)+ybar[i] # Replace each row
> gdot = gdot + diag(numeric(p)+ybar0)
> gdot = 100/ybar0^2 * gdot
> gdot

           [,1]       [,2]       [,3]       [,4]       [,5]
[1,] 877.26326 497.51642 497.51642 497.51642 497.51642
[2,] 165.83881 545.58564 165.83881 165.83881 165.83881
[3,] 305.23954 305.23954 684.98638 305.23954 305.23954
[4,]  33.64845  33.64845  33.64845 413.39529  33.64845
[5,]  60.08652  60.08652  60.08652  60.08652 439.83336
```

Asymptotic covariance matrix of percent differences

```
> # The approximate asymptotic covariance matrix of
> # percent differences  will be called V
>
> V = (1/n) * gdot %*% Sighat %*% t(gdot)
> carnames = c("Toyota", "Nissan", "Mazda", "Mitsubishi", "Suburu")
> rownames(V) = carnames; colnames(V) = carnames
> V

               Toyota    Nissan     Mazda Mitsubishi     Suburu
Toyota      95.777160 18.105819 33.325203  3.6736445 6.5600794
Nissan      18.105819 19.855174 11.108401  1.2245482 2.1866931
Mazda       33.325203 11.108401 45.882527  2.2538785 4.0247830
Mitsubishi   3.673644  1.224548  2.253878  3.0524969 0.4436769
Suburu       6.560079  2.186693  4.024783  0.4436769 5.7994905
```

Confidence intervals

```
> StdError = sqrt(diag(V))
> MarginError95 = 1.96*StdError
> PercentDiff = perdif[2:6]
> Lower95 = PercentDiff - MarginError95
> Upper95 = PercentDiff + MarginError95
>
> # Report
> rbind(car,Percent); rbind(PercentDiff,MarginError95,Lower95,Upper95)

        Honda Toyota Nissan  Mazda Mitsubishi Suburu
car    316.00  414.0  138.0 254.00      28.00  50.00
Percent 26.33   34.5   11.5  21.17       2.33   4.17


                 Toyota      Nissan       Mazda Mitsubishi      Suburu
PercentDiff    31.00000 -56.300000 -19.600000 -91.100000 -84.200000
MarginError95  19.18170   8.733592  13.276382   3.424394   4.720098
Lower95        11.81830 -65.033592 -32.876382 -94.524394 -88.920098
Upper95        50.18170 -47.566408  -6.323618 -87.675606 -79.479902
```

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 LaTeX source code is available from the course website:
http://www.utstat.toronto.edu/~brunner/oldclass/appliedf13