PLSC 502 – Autumn 2016 Measures of Association: Binary Variables

November 1, 2016

Binary Variables

- Ambiguous level of measurement...
- Related to proportions... For $Y \in \{0, 1\}$:
  - $E(Y) \equiv \sum Y/N = \hat{\pi}$
  - Same as $\widehat{\Pr(Y_i = 1)}$
  - Variance is $\hat{\pi}(1 - \hat{\pi})$
- Also potentially interval / ratio (as a "count")

Differences of Proportions

We know that for two estimates $\hat{\pi}_1$ and $\hat{\pi}_2$, based on samples of size N_1 and N_2,

$$z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\hat{\sigma}_{\pi_1 - \pi_2}}$$

where

$$\hat{\sigma}_{\pi_1 - \pi_2} = \sqrt{\frac{\hat{\pi}_1(1 - \hat{\pi}_1)}{N_1} + \frac{\hat{\pi}_2(1 - \hat{\pi}_2)}{N_2}}$$

We can think about this as samples of Y drawn from (say) $X = 0$ and $X = 1$:

$$\hat{\sigma}_{\pi_{Y|X=0} - \pi_{Y|X=1}} = \sqrt{\frac{\hat{\pi}_{Y|X=0}(1 - \hat{\pi}_{Y|X=0})}{N_{X=0}} + \frac{\hat{\pi}_{Y|X=1}(1 - \hat{\pi}_{Y|X=1})}{N_{X=1}}}$$

Chi-Square

We also know that:

$$W = \sum_{k_X k_Y} \frac{(N_{XY} - E_{XY})^2}{E_{XY}}$$

and that:

$$W \sim \chi_1^2$$

when both X and Y are binary.

In fact, $z^2 = W$...

```
> T <- table(Y,X)
> T
   X
Y   0 1
  0 5 3
  1 4 8

> chisq.test(T,correct=FALSE)

Pearson's Chi-squared test

data:  T
X-squared = 1.65, df = 1, p-value = 0.2

> p1<-4/9
> p2<-8/11
> p <- 12/20
> se <- sqrt(((p*(1-p)*(1/9+1/11))))
> Z <- (p1-p2) / se
> Z
[1] -1.2845
> Z^2
[1] 1.6498
```

χ^2 Is *Not* A Measure Of Association

```
> chisq.test(T, correct=FALSE)

Pearson's Chi-squared test

data:  T
X-squared = 1.65, df = 1, p-value = 0.199

> X <- rep(X,times=10)
> Y <- rep(Y,times=10)
> T10 <- table(X,Y)
> T10
   Y
X    0  1
  0 50 40
  1 30 80
> chisq.test(T10,correct=FALSE)

Pearson's Chi-squared test

data:  T10
X-squared = 16.5, df = 1, p-value = 0.0000487
```

"Contingency Tables"

Contingency table:

	$X = 0$	$X = 1$	
$Y = 0$	N_{00}	N_{10}	$N_{\bullet 0}$
$Y = 1$	N_{01}	N_{11}	$N_{\bullet 1}$
	$N_{0\bullet}$	$N_{1\bullet}$	N

Q: How much more or less likely is $Y = 1 | X = 1$ than $Y = 1 | X = 0$?

Odds

Recall that the *odds* of $Y = 1|X = 1$ are:

$$\begin{aligned} O_{Y=1|X=1} &= \frac{\Pr(Y = 1|X = 1)}{\Pr(Y = 0|X = 1)} \\ &= \frac{\hat{\pi}_{Y=1|X=1}}{\hat{\pi}_{Y=0|X=1}} \\ &= \frac{N_{11}/N_{1\bullet}}{N_{10}/N_{1\bullet}} \\ &= \frac{N_{11}}{N_{10}} \end{aligned}$$

And similarly:

$$O_{Y=1|X=0} = \frac{N_{01}}{N_{00}}$$

Odds Ratio

The *odds ratio* is then:

$$
\begin{aligned}
OR &= \frac{O_{Y=1|X=1}}{O_{Y=1|X=0}} \\
&= \frac{N_{11}/N_{10}}{N_{01}/N_{00}}
\end{aligned}
$$

Odds Ratio Facts...

- OR expresses the *relative* odds of an event ($Y = 1$) under one condition ($X = 1$) versus another ($X = 0$).
- $OR \in [0, \infty)$
- Interpretation:
  - $OR = 1 \;\leftrightarrow\;$ no association
  - $OR > 1 \;\leftrightarrow\;$ positive association
  - $OR < 1 \;\leftrightarrow\;$ negative association
- The "inverse odds ratio" $\left(O_{Y=0|X=1}/O_{Y=0|X=0}\right)$ is simply the reciprocal of OR.

Odds Ratios Illustrated

```
> T
   X
Y   0 1
  0 5 3
  1 4 8

> OR <- (T[1,1])*T[2,2] / (T[1,2]*T[2,1])
> OR
[1] 3.33333

> require(DescTools)
> OddsRatio(T)
[1] 3.33333
```

Association measure: ϕ

For the contingency table above,

$$\phi = \frac{N_{11}N_{00} - N_{10}N_{01}}{\sqrt{N_{1\bullet}N_{0\bullet}N_{\bullet 0}N_{\bullet 1}}}$$

Also,

$$\phi^2 = \frac{\chi^2}{N} \quad \text{so} \quad |\phi| = \sqrt{\frac{\chi^2}{N}}$$

A Few Things About ϕ

- A/K/A the "mean square contingency coefficient" or Matthews' Correlation Coefficient (MCC)
- $\phi \in [0, 1]$ (but see below...)
- In general:
  - $\phi \in [0.7, 1.0]$ = a strong positive association
  - $\phi \in [0.4, 0.7]$ = a moderate positive association
  - $\phi \in [0.1, 0.4]$ = a weak positive association
  - $\phi \in [-0.1, 0.1]$ = no association
  - $\phi \in [-0.1, -0.4]$ = a weak negative association
  - $\phi \in [-0.4, -0.7]$ = a moderate negatie association
  - $\phi \in [-0.7, -1.0]$ = a strong negative association
- ϕ equals Pearson's correlation coefficient (r) applied to two binary variables.
- The equation above means that $\phi^2 \times N \sim \chi^2_1$, which can be used for hypothesis testing (e.g., for $H_0 : \phi = 0$).

ϕ Examples...

```
> T
   X
Y   0 1
  0 5 3
  1 4 8

> require(psych)
> phi(T)
[1] 0.29

> cor(X,Y)
[1] 0.287213
```

ϕ Examples (continued)

```
> Tpos<-as.table(rbind(c(10,0),c(0,10)))
> phi(Tpos)
[1] 1

> Tneg<-as.table(rbind(c(0,10),c(10,0)))
> phi(Tneg)
[1] -1

> T0<-as.table(rbind(c(5,5),c(5,5)))
> phi(T0)
[1] 0
```

ϕ: Restricted Range

From the Stata manual (entry for `tetrachoric`):

from -1 to 1. To illustrate, consider the following set of tables for two binary variables, X and Z:

	Z = 0	Z = 1	
X = 0	$20 - a$	$10 + a$	30
X = 1	a	$10 - a$	10
	20	20	40

For a equal to 0, 1, 2, 5, 8, 9, and 10, the Pearson and tetrachoric correlations for the above table are

a	0	1	2	5	8	9	10
Pearson	0.577	0.462	0.346	0	-0.346	-0.462	-0.577
Tetrachoric	1.000	0.792	0.607	0	-0.607	-0.792	-1.000

Tetachoric Correlation (r_{tet})

Setup:

- N observations, with
- T_i a *latent* trait for each observation;
- two *raters*, $\{1, 2\}$, each of which
  - observes a "noisy" version of T_i:

$$
\begin{aligned}
T_i^{*1} &= T_i + e_{1i} \\
T_i^{*2} &= T_i + e_{2i}
\end{aligned}
$$

  - and gives a binary rating to i; equals 0 if $T_i < \tau$, 1 if $T_i > \tau$. Call these X_{1i} and X_{2i}.
- Assume that $\{e_{1i}, e_{2i}\} \sim \Phi_2(0, 0, 1, 1, \rho)$ (*bivariate normal*)

Digression: Bivariate Normals

The Bivariate Normal is:

$$\Pr(X_1, X_2) = \frac{1}{2\pi\sigma_{X_1}\sigma_{X_2}\sqrt{1-\rho^2}} \exp\left[\frac{-z}{2(1-\rho^2)}\right]$$

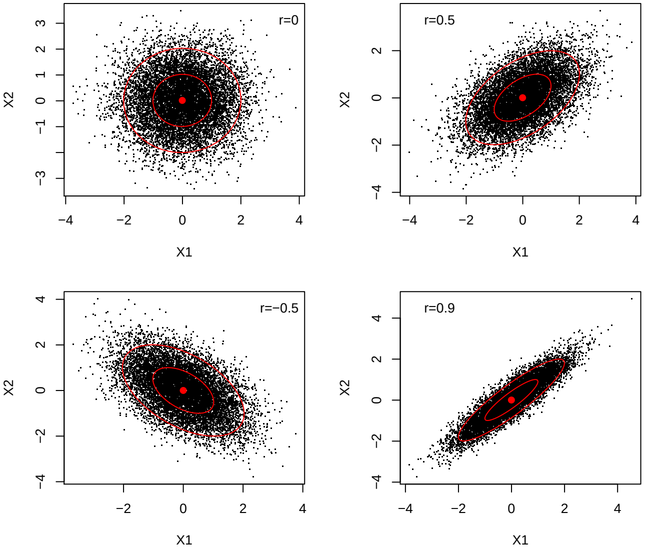
where

$$z = \left[\frac{(X_1 - \mu_{X_1})^2}{\sigma_{X_1}^2} + \frac{(X_2 - \mu_{X_2})^2}{\sigma_{X_2}^2} - \frac{2\rho(X_1 - \mu_{X_1})(X_2 - \mu_{X_2})}{\sigma_{X_1}\sigma_{X_2}}\right]$$

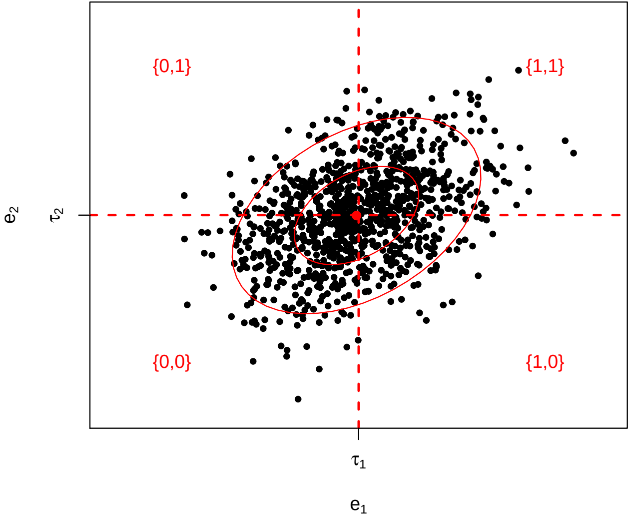
and

$$\rho \;=\; \text{corr}(X_1, X_2)$$

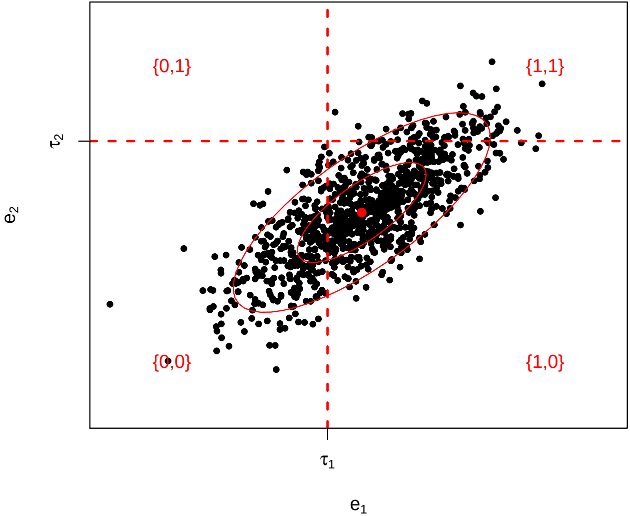
Bivariate Normals Illustrated



Back to Tetrachoric Correlation



Back to Tetrachoric Correlation



More Tetrachoric Correlation

Idea: Get as close to:

	$X_1 = 0$	$X_1 = 1$
$X_2 = 0$	π_{00}	π_{10}
$X_2 = 1$	π_{01}	π_{11}

...using three parameters: τ_1, τ_2, and ρ.

r_{tet} Fun Facts

- $r_{tet} \in [-1, 1]$
- Assumes two continuous, *Normal* underlying (latent) variables...
- Fitted via ML, etc. but also has a simple approximate formula:

$$r_{tet} \approx \frac{\alpha - 1}{\alpha + 1}$$

where

$$\alpha = (OR)^{\frac{\pi}{4}}$$

r_{tet}: An Example

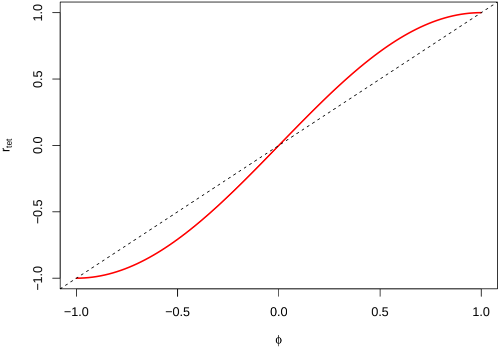
```
> require(polycor)
> polychor(T)
[1] 0.439917

> # Compare:
>
> phi(T)
[1] 0.29

> # Approximate formula:
>
> alpha <- (OR)^(pi/4)
> rtet <- (alpha - 1) / (alpha + 1)
> rtet
[1] 0.440458
```

r_{tet} vs. ϕ: Symmetrical Marginals

```
> addmargins(ST)
      A   B Sum
A     0 100 100
B   100   0 100
Sum 100 100 200
```



r_{tet} vs. ϕ: Asymmetrical Marginals

```
> addmargins(AT)
      A   B Sum
A     0 150 150
B   100 150 250
Sum 100 300 400
```