

2-by-2 Contingency Tables

Say we have n independent trials. For each trial we observe two binary variables X and Y . Let n_{ij} denote the number of trials where $X = i, Y = j$, for $i = 1, 2$ and $j = 1, 2$. This scenario is presented in the following table: Note that for the purposes of this discussion,

	$Y = 1$	$Y = 2$
$X = 1$	n_{11}	n_{12}
$X = 2$	n_{21}	n_{22}

we will use lower-case letters n_{ij} to represent random variables here, one could easily relabel this description with $n_{ij} \equiv N_{ij}$. We have that $n_{11} + n_{12} + n_{21} + n_{22} = n$.

We have that the probability distribution is specified by knowing n and this table:

	$Y = 1$	$Y = 2$
$X = 1$	p_{11}	p_{12}
$X = 2$	p_{21}	p_{22}

Where population parameters p_{ij} , satisfy $p_{11} + p_{12} + p_{21} + p_{22} = 1$.

To generate a single realization of a 2-by-2 contingency table, we can use the `simtab()` function that is defined in Dr. Shedden's lecture notes, where the function code is given below:

```
simtab = function(P, n)
{
  ## Convert to cumulative probabilities.
  CP = cumsum(P)
  ## Storage for the data being simulated.
  N = array(0, c(2,2))
  ## Simulate one contingency table.
  U = runif(n)
  N[1,1] = sum(U <= CP[1])
  N[1,2] = sum( (U > CP[1]) & (U <= CP[2]) )
  N[2,1] = sum( (U > CP[2]) & (U <= CP[3]) )
  N[2,2] = sum(U > CP[3])
  return(N)
}
```

This function takes two arguments, P is the vector of cell probabilities $(p_{11}, p_{12}, p_{21}, p_{22})'$ and n is the sample size. The function returns a 2-by-2 matrix N which is a single realization of the table with the specified parameters. An example function call:

```
n=1e4
P=(1:4)/sum(1:4)
N=simtab(P, n)
```

N is now the output matrix with the counts, we can convert these to proportions, by dividing entries in the table by the sample size n , and see if these proportions are in-line with the specified probabilities in P .

```
F=N/n
round(F, 3)
```

```
      [,1] [,2]
[1,] 0.105 0.199
[2,] 0.302 0.394
```

These are of course consistent with our specified cell probabilities. This function will be a useful tool for the homework problems which all require the generation of 2-by-2 contingency tables.

Another simple way to generate a single realization of a 2-by-2 contingency table is to use the `rmultinom()` function, for example:

```
N=t(array(rmultinom(n=1,size=n,prob=P),c(2,2)))
F=N/n
round(F, 3)
```

```
      [,1] [,2]
[1,] 0.104 0.193
[2,] 0.304 0.400
```

Odds/log-Odds Ratio

The odds of an event E , are $\frac{P(E)}{1-P(E)}$. For example the odds of seeing heads on a single fair coin flip is 1. The odds of rolling a 6 with a six sided die is 1/5. These are popularly read as 1 to 1 and 1 to 5 respectively.

Since probability is always in the range $[0, 1]$, we immediately have that the odds are in the range $[0, \infty]$, where infinite odds implies $P(E) = 1$.

The population odds ratio is the ratio of the odds that $Y = 1$ when $X = 1$ to the odds that $Y = 1$ when $X = 2$.

$$\text{OR} = \frac{p_{11}/p_{12}}{p_{21}/p_{22}} = \frac{p_{11}p_{22}}{p_{12}p_{21}}$$

Looking at the formula we notice that when the numerator is larger than the denominator, ($\text{OR} > 1$), then realizations (1, 1) and (2, 2) have higher probability. On the other hand,

when the denominator is larger than the numerator, ($OR < 1$), then realizations (1, 2) and (2, 1) have higher probability.

The population log-odds ratio (LOR) is the logarithm of the population odds ratio. The LOR can be any number in \mathbb{R} , where ($LOR > 0$) implies ($OR > 1$) and if ($LOR < 0$) then ($OR < 1$).

Sample Statistic	Population Parameter
Sample OR = $\frac{n_{11}n_{22}}{n_{12}n_{21}}$	OR = $\frac{p_{11}p_{22}}{p_{12}p_{21}}$
Sample LOR = $\log n_{11} + \log n_{22} - \log n_{12} - \log n_{21}$	LOR = $\log p_{11} + \log p_{22} - \log p_{12} - \log p_{21}$

The sample log-odds-ratio (SLOR):

$$SLOR = \log n_{11} + \log n_{22} - \log n_{12} - \log n_{21}$$

is used to estimate the population log-odds ratio. We are given that an approximate plug-in estimate of the standard error of this estimator is:

$$SE(SLOR) = \sqrt{1/n_{11} + 1/n_{12} + 1/n_{21} + 1/n_{22}}$$

For sufficiently large n we have an approximate result that a 95% CI for the LOR is:

$$SLOR \pm 2SE(SLOR)$$

We can use this interval to reject the null hypothesis of no association if this interval does not contain zero. This is a two-sided test of no association. If we realize a cell count of zero, we cannot reject the null hypothesis because the interval would include zero.

Recall that the power of a hypothesis test is the probability that it rejects the null hypothesis when it is in fact false. In problem 2 of the homework, you will generate data under the alternative hypothesis ($R \neq 0$), and compute the proportion of hypothesis tests that reject the null, using the above confidence interval. This proportion is the simulated estimate of the power.

Example

Examine the relationship between the sample size n and the power of the two-sided test of no association. Use population parameters $p_{11} = 0.1, p_{12} = 0.2, p_{21} = 0.3, p_{22} = 0.4$. Note that the population odds ratio here is $\frac{0.1 \cdot 0.4}{0.2 \cdot 0.3} = 2/3$, which is clearly less than 1 implying that there is association.

```
reps=1000
## Population Cell probs
P=(1:4)/sum(1:4)
```

```
POWER=NULL
for ( n in seq(from=10, to=5000, by=100) )
{
  ## Generate reps tables with sample size n
  N=rmultinom(n=reps,size=n,prob=P)

  ## Compute the sample log-odds ratios, and SEs
  SLOR = log(N[1,]*N[4,]/N[2,]/N[3,])
  SE = sqrt(1/N[1,] + 1/N[2,] + 1/N[3,] + 1/N[4,])

  ## Compute the 95% CI
  LB = SLOR-2*SE
  UB = SLOR+2*SE

  ## Get a list of indicators of
  ## CI's not including zero
  REJECT = (LB > 0 ) | (UB < 0)

  ## Do not reject if a cell count is zero
  REJECT[is.na(REJECT)]=FALSE
  POWER = c(POWER,mean(REJECT))
}

plot(seq(from=10, to=5000, by=100) , POWER, t="l",
      xlab="Sample Size,n", ylab="POWER", main="")
```

```
## The Same thing as above, but we use modified
## code from lecture notes
```

```
reps=1000
## Population Cell probs
P=(1:4)/sum(1:4)
```

```
POWER=NULL
for ( n in seq(from=10, to=5000, by=100) )
{
  ## Array of rejection indicators.
  REJECT = array(0, reps)
  for (k in 1:reps)
  {
```

```
## Generate a single table
N = simtab(P, n)
## The sample log-odds ratio and its standard error.
LR = log(N[1,1]) + log(N[2,2]) - log(N[1,2]) - log(N[2,1])
SE = sqrt(1/N[1,1] + 1/N[1,2] + 1/N[2,1] + 1/N[2,2])

## The 95 Percent CI
LB = LR - 2*SE
UB = LR + 2*SE

## Check to see if Zero is outside the interval.
if (!is.finite(SE)) { REJECT[k] = 0 }
else { REJECT[k] = (LB > 0) | (UB < 0) }
}
POWER=c(POWER, mean(REJECT))
}

plot(seq(from=10, to=5000, by=100), POWER, t="l",
      xlab="Sample Size,n", ylab="POWER", main="")
```