STAT 311: More Two - Way Tables

Y. Samuel Wang

Summer 2016

Logistics

- Midterm is Jul 22 (in class)
- Review in lab section Jul 21
- Homework due now
- Lab due Monday at midnight

Two Way Table Review

- Two way tables summarize bivariate categorical data
- Select a row variable and a column variable
- Each cell represents the counts of individuals who satisfy both the row and column characteristics

Notation

- Marginal Distribution / Probability of A: $P(A)$
- Joint Distribution / Probability of A **and** B: $P(A, B)$ or $P(A \cap B)$
- Conditional Distribution / Probability of A **given** B: $P(A|B)$

Two Way Table Review

Consider the Ebola data from yesterday's lab[1]

	Death	Survive	Total
Guinea	2536	1268	3804
Liberia	4806	5860	10666
Sierra Leone	3955	10167	14122
Total	11297	17295	28592

[1] Data available from World Health Organization: `http://apps.who.int/gho/data/view.ebola-sitrep.ebola-summary-latest?lang=en`. Up to date as of Dec 2015

Joint Distribution

Divide all inner cells by the grand total to get the **joint distribution**. This describes the proportion of all observational units which fall into *both* the row and column categories

	Death	Survive
Guinea	0.09	0.04
Liberia	0.17	0.20
Sierra Leone	0.14	0.36

Joint Distribution

Divide all inner cells by the grand total to get the **joint distribution**. This describes the proportion of all observational units which fall into *both* the row and column categories

	Death	Survive
Guinea	0.09	0.04
Liberia	0.17	0.20
Sierra Leone	0.14	0.36

Joint distribution answers what proportion of all individuals satisfy both R_i and C_j?

Joint Distribution

Divide all inner cells by the grand total to get the **joint distribution**. This describes the proportion of all observational units which fall into *both* the row and column categories

	Death	Survive	Total
Guinea	2536	1268	3804
Liberia	4806	5860	10666
Sierra Leone	3955	10167	14122
Total	11297	17295	28592

$$P(\text{Guinea}, \text{Death}) = P(\text{Guinea} \cap \text{Death}) = \frac{2536}{28592}$$

Marginal Distribution

Divide all row (or column) totals by the grand total to get the **marginal distribution**. This describes the proportion of all observational units which fall into *both* each of the row (or column) categories.

Death	Survive
0.40	0.60

Guinea	Liberia	Sierra Leone
0.20	0.26	0.54

Marginal Distribution

Divide all row (or column) totals by the grand total to get the **marginal distribution**. This describes the proportion of all observational units which fall into *both* each of the row (or column) categories.

Death	Survive
0.40	0.60

Guinea	Liberia	Sierra Leone
0.20	0.26	0.54

Marginal distribution answers what proportion of all individuals are in category R_i? What proportion of all individuals are C_i?

Marginal Distribution

Divide all row (or column) totals by the grand total to get the **marginal distribution**. This describes the proportion of all observational units which fall into *both* each of the row (or column) categories.

	Death	Survive	Total
Guinea	2536	1268	3804
Liberia	4806	5860	10666
Sierra Leone	3955	10167	14122
Total	11297	17295	28592

$$P(\text{Guinea}) = \frac{3804}{28592}$$

Conditional Distribution

Divide all inner cells by the row (or column) total to get the **conditional distribution** conditioned on the row (or column). This describes the proportion of all observational units in a row (or column) which fall into each of the column (or row) categories.

Table: Conditioning on Row

	Death	Survive	Total
Guinea	0.667	0.333	1.00
Liberia	0.451	0.549	1.00
Sierra Leone	0.280	0.720	1.00

Table: Conditioning on Column

	Death	Survive
Guinea	0.224	0.073
Liberia	0.425	0.339
Sierra Leone	0.350	0.588
Total	1.00	1.00

Conditional Distribution

Divide all inner cells by the row (or column) total to get the **conditional distribution** conditioned on the row (or column). This describes the proportion of all observational units in a row (or column) which fall into each of the column (or row) categories.

	Death	Survive	Total
Guinea	2536	1268	3804
Liberia	4806	5860	10666
Sierra Leone	3955	10167	14122
Total	11297	17295	28592

Given that a case is in Liberia, what is the probability of death?

$$P(Death|Liberia) = \frac{4806}{10666}$$

Conditional Distribution

Divide all inner cells by the row (or column) total to get the **conditional distribution** conditioned on the row (or column). This describes the proportion of all observational units in a row (or column) which fall into each of the column (or row) categories.

	Death	Survive	Total
Guinea	2536	1268	3804
Liberia	4806	5860	10666
Sierra Leone	3955	10167	14122
Total	11297	17295	28592

Given that a case resulted in Death, what is the probability that it is Liberia?

$$P(Liberia|Death) = \frac{4806}{11297}$$

Risk Ratios

The probability of an event can also be referred to as the "risk" regardless of whether the outcome is good or bad. We can compute the **Relative Risk** by looking at a ratio of the conditional distributions of two different categories.

$$\text{Relative Risk of outcome A between groups} B_1 \text{ and } B_2 = \frac{P(A|B_1)}{P(A|B_2)}$$

Risk Ratios

The probability of an event can also be referred to as the "risk" regardless of whether the outcome is good or bad. We can compute the **Relative Risk** by looking at a ratio of the conditional distributions of two different categories.

$$\text{Relative Risk of outcome A between groups} B_1 \text{ and } B_2 = \frac{P(A|B_1)}{P(A|B_2)}$$

Relative Risk answers the question "How much more likely is an event to occur given that B_1 occurred instead of B_2?

Risk Ratios

Table: Conditioning on Row

	Death	Survive	Total
Guinea	0.667	0.333	1.00
Liberia	0.451	0.549	1.00
Sierra Leone	0.280	0.720	1.00

What is the relative risk between Guinea and Liberia of an ebola case resulting in death?

$$\frac{\text{Risk of dying, given that it occured in Guinea}}{\text{Risk of dying, given that it occured in Liberia}} = \frac{P(Death|Guinea)}{P(Death|Liberia)}$$

$$= \frac{.667}{.451} = 1.48$$

Risk Ratios

Table: Conditioning on Row

	Death	Survive	Total
Guinea	0.667	0.333	1.00
Liberia	0.451	0.549	1.00
Sierra Leone	0.280	0.720	1.00

What is the relative risk between Guinea and Liberia of an ebola case resulting in death?

$$\frac{\text{Risk of dying, given that it occured in Guinea}}{\text{Risk of dying, given that it occured in Liberia}} = \frac{P(Death|Guinea)}{P(Death|Liberia)}$$

$$= \frac{.667}{.451} = 1.48$$

A case of ebola is 1.48 times more likely to result in death in Guinea than it is in Liberia

Odds

- Odds colloquially mean just the probability, but in statistics they have a more technical definition
- The **odds** of an event is the ratio of the probability that the event occurs to the probability that the event does not occur

$$\text{odds} = \frac{P(A)}{P(\text{not A})}$$

- Odds are sometimes communicated as "probability A occurs" to "probability A does not occur" odds. 3 to 2 odds implies that the probability of an event happening is 3/2 times more likely than the event not happening

Odds

- Odds colloquially mean just the probability, but in statistics they have a more technical definition
- The **odds** of an event is the ratio of the probability that the event occurs to the probability that the event does not occur

$$\text{odds} = \frac{P(A)}{P(\text{not A})}$$

- Odds are sometimes communicated as "probability A occurs" to "probability A does not occur" odds. 3 to 2 odds implies that the probability of an event happening is 3/2 times more likely than the event not happening

Odds Ratios

Just like we can take a risk ratio, we can also take a ratio of odds for two separate groups.

Table: Conditioning on Row

	Death	Survive	Total
Guinea	0.667	0.333	1.00
Liberia	0.451	0.549	1.00
Sierra Leone	0.280	0.720	1.00

The odds of a death given that a case arose in Guinea is $\frac{.667}{.333}$. The odds of a death given that a case arose in Sierra Leone is $\frac{.28}{.72}$.

Odds Ratios

Just like we can take a risk ratio, we can also take a ratio of odds for two separate groups.

Table: Conditioning on Row

	Death	Survive	Total
Guinea	0.667	0.333	1.00
Liberia	0.451	0.549	1.00
Sierra Leone	0.280	0.720	1.00

The odds of a death given that a case arose in Guinea is $\frac{.667}{.333}$. The odds of a death given that a case arose in Sierra Leone is $\frac{.28}{.72}$.

Thus, the odds ratio of death in Guinea vs Sierra Leone is $\frac{.667/.333}{.28/.72}$

Notes about odds vs risk

- If the odds of an event is k, then the risk (or probability) of that event is $k/(1+k)$
- If the risk (or probability) of an event is p, the odds of that event is $p/(1-p)$

Notes about odds vs risk

- If the odds of an event is k, then the risk (or probability) of that event is $k/(1+k)$
- If the risk (or probability) of an event is p, the odds of that event is $p/(1-p)$

- Risk ratios are easier to interpret
- Odds ratios are still used because the math can be easier in certain situations

Caution about ratios

When we take the ratio of risks (or odds), it can be useful in comparing two groups, but it can also be misleading in some cases.

- 2015 study showed that eating bacon increased the risk of colorectal cancer by 18%
- The risk ratio of colorectal cancer between those who ate bacon regularly and those who did not is 1.18
- However, in absolute terms, this means that the chance of colorectal cancer increased from 5% to 6%[2]

[2] LA times article on how to interpret the study: `http://www.latimes.com/opinion/opinion-la/la-ol-bacon-risk-20151030-story.html`

Caution about ratios

When we take the ratio of risks (or odds), it can be useful in comparing two groups, but it can also be misleading in some cases.

- 2015 study showed that eating bacon increased the risk of colorectal cancer by 18%
- The risk ratio of colorectal cancer between those who ate bacon regularly and those who did not is 1.18
- However, in absolute terms, this means that the chance of colorectal cancer increased from 5% to 6%[2]

Statistical significance, does not always mean practical significance

[2] LA times article on how to interpret the study: `http://www.latimes.com/opinion/opinion-la/la-ol-bacon-risk-20151030-story.html`

Simpson's Paradox

When examining two variables, there might actually be a third **confounding variable**, which explains the perceived association between the other two variables. When the confounding variable flips the direction of the association, it is called **Simpson's paradox**.

Simpson's Paradox

Suppose we have data from two hospitals, which hospital seems to be doing a better job?

Table: All patient outcomes

	Survive	Death
Hospital A	500	500
Hospital B	200	310

Simpson's Paradox

But let's take a deeper dive in the data. Now which hospital seems to be doing a better job?

Table: Patients with paper cuts

	Survive	Death
Hospital A	400	100
Hospital B	50	10

Table: Patients bitten by zombies

	Survive	Death
Hospital A	100	400
Hospital B	150	300

Simpson's Paradox

But let's take a deeper dive in the data. Now which hospital seems to be doing a better job?

Table: Patients with paper cuts

	Survive	Death
Hospital A	400	100
Hospital B	50	10

Table: Patients bitten by zombies

	Survive	Death
Hospital A	100	400
Hospital B	150	300

The aggregated data hides the fact that Hospital A sees a relatively larger portion of patients with less serious injuries than Hospital B. The seriousness of injuries would be a confounding variable.

Sensitivity vs Specificity

One specific use of two-way tables involves medical tests

	Truly Sick	Truly Well	Total
Diagnosed as Sick	300	50	350
Diagnosed as Well	100	500	600
Total	400	550	950

Sensitivity vs Specificity

One specific use of two-way tables involves medical tests

	Truly Sick	Truly Well	Total
Diagnosed as Sick	300	50	350
Diagnosed as Well	100	500	600
Total	400	550	950

If the test determines that the patient is sick, we say they have tested positive

	Truly Sick	Truly Well	Total
Diagnosed as Sick	True Positive	False Positive	Positive Tests
Diagnosed as Well	False Negative	True Negative	Negative Tests
Total	Total Sick	Total Well	Total Tested

Sensitivity vs Specificity

Conditioning on the true status

- **Sensitivity** is the probability that some who is truly positive, tests positive. $P(Test+|True+)$
- **Specificity** is the probability that someone who is truly negative, tests negative. $P(Test-|True-)$

Sensitivity vs Specificity

Conditioning on the true status

- **Sensitivity** is the probability that some who is truly positive, tests positive. $P(Test+|True+)$
- **Specificity** is the probability that someone who is truly negative, tests negative. $P(Test-|True-)$

Conditioning on the test outcome

- **Positive Predictive Value** is the probability that some who is tests positive, is truly positive. $P(True+|Test+)$
- **Negative Predictive Value** is the probability that someone who is truly negative, tests negative. $P(True-|Test-)$

Sensitivity vs Specificity

	Truly Sick	Truly Well	Total
Diagnosed as Sick	300	50	350
Diagnosed as Well	100	500	600
Total	400	550	950

- Specificity $P(Test - | True-)$:

Sensitivity vs Specificity

	Truly Sick	Truly Well	Total
Diagnosed as Sick	300	50	350
Diagnosed as Well	100	500	600
Total	400	550	950

- Specificity $P(Test - | True-)$: 500/550
- Sensitivity $P(Test + | True+)$:

Sensitivity vs Specificity

	Truly Sick	Truly Well	Total
Diagnosed as Sick	300	50	350
Diagnosed as Well	100	500	600
Total	400	550	950

- Specificity $P(Test - | True-)$: $500/550$
- Sensitivity $P(Test + | True+)$: $300/400$
- PPV $P(True + | Test+)$:

Sensitivity vs Specificity

	Truly Sick	Truly Well	Total
Diagnosed as Sick	300	50	350
Diagnosed as Well	100	500	600
Total	400	550	950

- Specificity $P(Test-|True-)$: 500/550
- Sensitivity $P(Test+|True+)$: 300/400
- PPV $P(True+|Test+)$: 300/350
- NPV $P(True-|Test-)$:

Sensitivity vs Specificity

	Truly Sick	Truly Well	Total
Diagnosed as Sick	300	50	350
Diagnosed as Well	100	500	600
Total	400	550	950

- Specificity $P(Test-|True-)$: $500/550$
- Sensitivity $P(Test+|True+)$: $300/400$
- PPV $P(True+|Test+)$: $300/350$
- NPV $P(True-|Test-)$: $500/600$