STATS 111/202

Lecture 3: sampling, more tables, and ordinal data

1/13/2025

Ana Maria Kenney

UC Irvine Department of Statistics

Notation reminder

- In the textbook, when considering proportions and conditional probabilities the author uses the notation:
  - π – population parameter
  - p – sample estimate of population parameter

- In previous classes we often see:
  - p – population parameter
  - $\hat{p}$ - sample estimate of population parameter

- For this homework, we will accept either format. Just make sure you are differentiating between population vs sample estimate

- For future lectures, we will stick with “hat” notation. This will likely match up better with future courses and previous courses

Refresh of study/sampling types

2.1.5, 2.3.6

Studies

- **Observational study**
    - Participants are only observed or surveyed and measured
    - When comparisons between groups are made, these are **naturally** formed
    - **Cause and effect conclusions cannot generally be made**

- **Randomized experiment**
    - Participants are randomly assigned to treatment vs control groups
    - Should **account for potential confounding variables** across groups
    - **When well-designed, cause and effect conclusions can generally be made**

Studies

- **Confounding variables**
  - A variable that affects the response but is also related to the explanatory variable
  - In this case, its effect on the response cannot be separated from the effect on the explanatory variable
  - **Example: eating more ice cream can lead to shark attacks**

Designs

- **Prospective**
  - Collect the sample, then follow participants for a length of time to observe future events during the study period
  - **Examples:**
    - **Cohort study: an observational study in which a cohort (a group sharing a common characteristic) is followed throughout the study and outcome variables are observed.**
    - **Clinical trial: a randomized experiment where participants are randomly assigned treatments then followed throughout the study as outcome variables are observed**
- **Retrospective**
  - Collect the sample, then ask about the past
  - **Example: case-control study – observational study where a sample of cases is collected along with controls then explanatory variables are compared.**

Designs

- **Cross-sectional**
  - A representative sample of a population is collected, and data are collected on those individuals at a specific point in time
  - This is an observational study, but neither prospective or retrospective

Two-way tables

2.1-2.4

Two-way tables

General example

	Y=1	Y=0	Total
X=1	n_{11}	n_{12}	$n_{1+} = n_{11} + n_{12}$
X=0	n_{21}	n_{22}	$n_{2+} = n_{21} + n_{22}$
Total	$n_{+1} = n_{11} + n_{21}$	$n_{+2} = n_{12} + n_{22}$	n

Consider the scenario where a fixed number of subjects with X=0 and X=1 are sampled, and then we observe to see if they result in a Y=0 or Y=1

- Here n_{1+} and n_{2+} are fixed
- Take $p_0 = P(Y = 1|X = 0)$ and $p_1 = P(Y = 1|X = 1)$
- Then $n_{11} \sim Binomial(n_{1+}, p_1)$ and $n_{21} \sim Binomial(n_{2+}, p_0)$
- We compare p_0 to p_1

Two-way tables

Other important quantities:

- Risk difference $RD = p_1 - p_0$
- Relative risk (risk ratio): $RR = \frac{p_1}{p_0}$
- Odds ratio OR$= \frac{p_1/(1-p_1)}{p_0/(1-p_0)}$
- Odds is the probability of an event divided by the probability of no event

Two-way tables

Odds ratio (OR)

- Strictly nonnegative ($0 \leq OR < \infty$)
- When p_0 and p_1 are close to 0, OR will approximate RR
  - Because $(1 - p_1)$ and $(1 - p_0)$ are approximately equal to 1

Two-way tables

Other important quantities:

- Risk difference $RD = p_1 - p_0$

Estimate: $\hat{p}_1 - \hat{p}_0$

- Relative risk (risk ratio): $RR = \frac{p_1}{p_0}$

Estimate: $\frac{\hat{p}_1}{\hat{p}_0}$

- Odds ratio OR= $\frac{p_1/(1-p_1)}{p_0/(1-p_0)}$

Estimate: $\frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_0/(1-\hat{p}_0)}$

- Odds is the probability of an event divided by the probability of no event

Two-way tables

Estimates using table quantities

- $\hat{p}_1 = \frac{n_{11}}{n_{1+}}$
- $\hat{p}_0 = \frac{n_{21}}{n_{2+}}$
- $\widehat{RD} = \hat{p}_1 - \hat{p}_0$
- $\widehat{RR} = \frac{\hat{p}_1}{\hat{p}_0}$
- $\widehat{OR} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_0/(1-\hat{p}_0)}$

	Y=1	Y=0	Total
X=1	n_{11}	n_{12}	$n_{1+} = n_{11} + n_{12}$
X=0	n_{21}	n_{22}	$n_{2+} = n_{21} + n_{22}$
Total	$n_{+1} = n_{11} + n_{21}$	$n_{+2} = n_{12} + n_{22}$	n

Inference

Using some careful math, we can establish asymptotic normality of the estimates and compute standard errors:

Standard error estimates	$100(1-\alpha)\%$ CI
$\widehat{se}(\widehat{RD}) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_{1+}} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_{2+}}}$	$\hat{p}_1 - \hat{p}_0 \pm z_{\alpha/2} \times \widehat{se}(\widehat{RD})$
$\widehat{se}(\log(\widehat{RR})) = \sqrt{\frac{(1-\hat{p}_1)}{\hat{p}_1 n_{1+}} + \frac{(1-\hat{p}_0)}{\hat{p}_0 n_{2+}}}$	$\log(\frac{\hat{p}_1}{\hat{p}_0}) \pm z_{\alpha/2} \times \widehat{se}(\log(\widehat{RR}))$
$\widehat{se}(\log(\widehat{OR})) = \sqrt{\frac{1}{\hat{p}_1 n_{1+}} + \frac{1}{(1-\hat{p}_1)n_{1+}} + \frac{1}{\hat{p}_0 n_{2+}} + \frac{1}{(1-\hat{p}_0)n_{2+}}}$	$\log(\frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_0/(1-\hat{p}_0)}) \pm z_{\alpha/2} \times \widehat{se}(\log(\widehat{OR}))$

Inference

In general, for an estimate $\widehat{\boldsymbol{\theta}}$ of population parameter $\boldsymbol{\theta}$, the 100$(1-\alpha)$% CI follows the form:

$$(I, u) = (\hat{\theta} - z_{\alpha/2} se(\hat{\theta}), \hat{\theta} + z_{\alpha/2} se(\hat{\theta}))$$

Then we can get a CI for the log transformed RR and OR by back transformation (exponentiate it)

$$(RR_l, RR_u) = (e^l, e^u)$$

Inference: hypothesis testing

- As mentioned last week, generally interested in whether $p_1 = p_0$
- Meaning, RD=0 or RR=1 or OR=1
- If true, then the categorical explanatory variable has no effect on the outcome!
- Testing RD=0 is equivalent to testing if RR=OR=1
- The null H_0 follows RD=0 or RR=OR=1, and the alternative H_α for RD can be of the form $\neq 0$ or one-sided $< 0\ or > 0$ (and $\neq 1$ or one-sided $< 1\ or > 1$ for RR or OR)

CHD example #1

IxJ (I by J) tables

- When the number of categories of X and/or Y increase, we will have many pair-wise comparisons between levels of X and levels of Y

- Let the inference goal be whether **the level of X** is associated with the **level of Y**

- Let the explanatory variable X have I many levels, i=1,2,…,I

- Let the response Y have J levels, j=1,2,…,J

IxJ (I by J) tables

	Y=1	**Y=2**	$\cdots$	**Y=J**	**Total**
X=1	n_{11}	n_{12}	$\cdots$	n_{1J}	n_{1+}
X=2	n_{21}	n_{22}	$\cdots$	n_{2J}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
X=I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}	n_{I+}
Total	n_{+1}	n_{+2}	$\cdots$	n_{+J}	n

IxJ (I by J) tables

- Now we want to test whether X and Y are independent (is X associated with Y?)

- X and Y are independent if the conditional distributions of Y given X are the same at each level of X

- Meaning, the **level of X has no effect on the distribution of Y**

IxJ (I by J) tables

- Let $p_{ij} = P(X = i, Y = j)$
- X and Y are independent if and only if $p_{ij} = p_{i+}p_{+j}$
- Where $p_{i+} = P(X = i)$ and $p_{+j} = P(Y = j)$ for i=1,2,…,I and j=1,2,…,J

IxJ (I by J) tables

- The null hypothesis will then be: $H_0: p_{ij} = p_{i+}p_{+j}\ for\ all\ i = 1,2,\dots,I\ and\ j = 1,2,\dots,J$

- The alternative will then be: $H_\alpha: at\ least\ one\ combination\ of\ i\ and\ j\ has\ p_{ij} \neq p_i p_j$

- The test statistics is of the form $\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(n_{ij}-\hat{\mu}_{ij}\right)^2}{\hat{\mu}_{ij}}$

- Where $\hat{\mu}_{ij}$ is the **estimated expected number** of observations where X=i, Y=j, **under the null**
    - The expected number of n_{ij} where n_{ij} is the observed number of observations

Computing expected number of observations in a given cell

- Estimate μ_{ij} by
  - $\hat{\mu}_{ij} = n\,\hat{p}_{ij} \; = n\hat{p}_{i+}\hat{p}_{+j} = n\frac{n_{i+}}{n}\frac{n_{+j}}{n} = \frac{n_{i+}n_{+j}}{n}$

Test statistic

- From our test statistic, we have a chi-squared distribution with (I-1)(J-1) degrees of freedom

- To obtain a p-value, we compute the area above the test statistic using a chi-squared distribution with (I-1)(J-1) degrees of freedom

  - In R, this would be 1-pchisq(X,(I-1)(J-1)), where X is the test statistic that was computed

Degrees of freedom

- Like before in regression, the degrees of freedom is the difference in the number of parameters between the full model and the reduced model

- The reduced model is the one under the null. So we have I-1 parameters to be estimated for the p_{+i} and another J-1 for the p_{j+}

- This gives a total of (I-1) + (J-1) parameters **under the null**

- The unconstrained model has to estimate all p_{ij} of which there are IJ-1 to estimate

- So the difference and total df follows: (IJ-1) – [(I-1) + (J-1)] = (I-1)(J-1)

Testing for independence in R

- We use the chisq.test() function
- Need to input the entire IxJ table
- The null is that X and Y are independent and the alternative is that this does not hold (i.e., X and Y are dependent)
- Can also have R compute all expected counts under the null
  - Say your IxJ table is stored under object ‘x’
  - To get the expected counts take: chisq.test(x)$expected

CHD example #2

Ordinal data

2.5

Incorporating ordinal structure

- We will extend these methods to account for explanatory variables that are ordinal in nature
    - Ex: medicine dosage, blood pressure
    - It can be argued that it is reasonable to assume the probability of the response being 1 (Y=1) is increasing (or decreasing) as we move down the two-way table

- Incorporating this information allows for testing more precise alternatives

- It is a stronger scientific statement if one hypothesizes a linear trend and later rejects in favor of that alternative

Incorporating ordinal structure

- Now we assign numeric scores to each possible outcome

  - Scores $u_1 \leq u_2 \leq \cdots \leq u_I$ for the I levels of X
  - Scores $v_1 \leq v_2 \leq \cdots \leq v_J$ for the J levels of Y

- Scores should match the ordering believed to be inherent in the categorical levels.

- Categories close to one another should have scores close together and vice versa

- A common approach is to naturally order outcomes (e.g., $u_1 = 1, u_2 = 2, \ldots, u_I = I$ and the same for v scores)

Incorporating ordinal structure

- With scores assigned, this is closer to our regression setting. We can look for a linear trend and use sample correlation between scores to form a test statistic

- Recall correlation: $\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$

- Under the scores we have:
  - $r = \frac{\sum_{i,j}(u_i - \bar{u})(v_j - \bar{v})p_{ij}}{\sqrt{[\sum_i (u_j - \bar{u})^2 p_{i+}][\sum_j (v_j - \bar{v})^2 p_{+j}]}}$
  - Where $\bar{u} = \sum_i u_i p_{i+}$ denotes the sample mean of row scores
  - $\bar{v} = \sum v_j p_{+j}$ denotes the sample mean of column scores

Incorporating ordinal structure

- We use the test statistic: $M^2 = (n-1)r^2 \sim \chi_1^2$
  - Meaning, it is distributed approximately chi-squared with 1 df
- Since we have 1 df, this means the full model has 1 extra parameter than the reduced
- Similar to testing a single slope parameter in a linear regression setting where the explanatory variable is the categorical explanatory variable being treated like a discrete ordinal variable (or rather, continuous)

Incorporating ordinal structure

- Let's say we have two levels for Y (0/1) and I levels for X

- Let $p_j = P(Y = 1|X = j)$, the probability Y=1 given the jth category of X

- For hypothesis testing, we have
  - $H_o: p_1 = p_2 = \cdots = p_i$ (X and Y are independent)
  - $H_\alpha: p_1 < p_2 < \cdots < p_I \ or \ p_1 > p_2 > \cdots > p_I$ (testing a trend, either increasing or decreasing)

- This assumes scores were generated such that the are increasing

- This is a trend test, and will give evidence that there is a trend in probabilities or not

  - Need to look at data to see whether it is increasing or decreasing

CHD example #3

Announcements

- HW 1 due this Sunday at 11:59pm PT
- For additional practice:
  - See questions 1.1, 1.3, 1.7, 1.9, 1.15, 2.1a, 2.3, 2.5, 2.9, 2.11, 2.13, 2.15, 2.19a, 2.21 in the textbook (all have solutions in the back/online)