Categorical data methods: Outline

- inference for one proportion
- inference for two proportions
- chi-squared tests (multinomial, goodness-of-fit)
- paired proportions

Inference for a single proportion

- Assume independent n identical trials, $Y_i,\ i = 1 \dots n$, binary (zero or one) responses, with constant $\text{Pr}(\text{success}) = \pi$
- define $Y = \sum_{i=1}^{n} Y_i = \#$ of successes in n trials
- define $p = \frac{Y}{n} =$ sample proportion of successes
- we write $Y \sim \text{Bin}(n, \pi)$
    - $f(y) = \text{Pr}(Y = y) = \binom{n}{k} \pi^k (1 - \pi)^{n-k}$ for $y = 0, 1, \dots, n$
    - $EY = n\pi$
    - $\text{Var}(Y) = n\pi(1 - \pi)$
    - note $p = Y/n$ is not binomial; $Ep = \pi$ and $\text{Var}(p) = \pi(1 - \pi)/n$

Inference for a single proportion

- Independence of individual events (0/1 responses) is crucial!
- A corollary of independence:
  each trial has same π
- Violation of either $\rightarrow$ wrong Var p
- Key result:
  - if $n\pi \geq 5$ and $n(1-\pi) \geq 5$,
    then p is approx $N(\pi, \pi(1-\pi)/n)$
  - Approximate $100(1-\alpha)\%$ CI for π is

$$p \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}},$$

    can be < 0 or > 1,
  - better approx. available (Fleiss, Stat. Meth. for Rates and Proportions).

Inference for a single proportion

- Test $H_o : \pi = \pi_o$ using test statistic

$$z = \frac{p - \pi_o}{\sqrt{\pi_o(1 - \pi_o)/n}}$$

  P-value from standard normal distn
- Note: variance calculated at π_o (H_o value of π)
- If sample size is too small for above test or CI, then use exact binomial calculations (i.e. a randomization test)

Inference for two proportions

- Now consider methods for two proportions
- $Y_1 \sim \text{Bin}(n_1, \pi_1)$ and $Y_2 \sim \text{Bin}(n_2, \pi_2)$
  Y_1 and Y_2 are independent r.v's.
- Goal (for now) is inference for $\pi_1 - \pi_2$
- Assume n_1 and n_2 are sufficiently large (usual rule)
- Basic result:
  - $p_1 - p_2$ is approx $N\left(\pi_1 - \pi_2, \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right)$
  - $100(1-\alpha)\%$ confidence interval

$$p_1 - p_2 \pm z_{1-\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Inference for two proportions

- Test $H_o : \pi_1 = \pi_2$
  - note: under H_o std error of $p_1 - p_2$ is different than given on previous slide (since $\pi_1 = \pi_2$)
  - use pooled estimate $p = (Y_1 + Y_2)/(n_1 + n_2)$
  - $z = (p_1 - p_2)/\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$
  - P-value from standard normal distn

Odds and Odds ratio

- Definition: odds in favor of success $= \pi_1/(1-\pi_1)$
- Odds ratio for pop'n 2 relative to pop'n 1

$$\phi = \frac{\pi_2/(1-\pi_2)}{\pi_1/(1-\pi_1)} = \frac{\pi_2(1-\pi_1)}{(1-\pi_2)\pi_1}$$

- Interpretation:
  - $\phi = 1$ means no difference in odds/proportions
    $\phi > 1$ means event more likely in population 2.
  - shows up frequently in medical statistics
  - later models allow for multiplicative changes to odds (e.g., logistic regression)
  - $\log \phi$ commonly used, is symmetric around 0

Inference for odds ratio

- Estimate: $\hat{\phi} = \frac{p_2\,(1-p_1)}{(1-p_2)\,p_1} = \frac{Y_2\,(n_1-Y_1)}{(n_2-Y_2)\,Y_1}$
- For large n,
  $\log\hat{\phi} \sim N(\log\phi, \frac{1}{n_1\pi_1(1-\pi_1)} + \frac{1}{n_2\pi_2(1-\pi_2)})$
- $\hat{\text{Var}}\ \log\hat{\phi} = \frac{1}{Y_1} + \frac{1}{n1-Y1} + \frac{1}{Y_2} + \frac{1}{n2-Y2}$
- If 0's, add 0.5 to all counts:

$$\log\hat{\phi} = \frac{(Y_2+0.5)\,(n_1-Y_1+0.5)}{(n_2-Y_2+0.5)\,(Y_1+0.5)}$$

$$\hat{\text{Var}}\ \log\hat{\phi} = \frac{1}{Y_1+0.5} + \frac{1}{n1-Y1+0.5} + \frac{1}{Y_2+0.5} + \frac{1}{n2-Y2+0.5}$$

Contingency tables

- Categorical data is often recorded in contingency tables

Rows	Columns 1	2	$\cdots$	c
1	n_{11}	n_{12}	$\cdots$	n_{1c}
2	n_{21}	n_{22}	$\cdots$	n_{2c}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	n_{r1}	n_{r2}	$\vdots$	n_{rc}

- Also called cross-classification tables or $r \times c$ table
- Can also be more than two-dimensional (we don't consider higher-dimensions here)
- We assume r rows and c columns

Contingency tables

- Examples
  - comparing two proportions (2×2 table with rows = populations, cols = success/failure)
  - comparing more than two proportions ($r \times 2$ table)
  - comparing two multinomial distns (more than two outcomes for each of two populations in a $2 \times c$ table)
  - comparing more than two multinomial distns ($r \times c$ table)
  - analyzing a single population classified on two dimensions (test for indep of the two dimens)
  - also allow possibility of $1 \times c$ table (test for goodness-of-fit to model)

Contingency tables

- Three possible probability structures for the counts in the table
  - a) If each row is a different population then it is natural to think of the proportions in each row ($\pi_{ij},\ j = 1, \ldots, c$) as summing to one
  - If the table is a single population then it is natural to think of the proportions in the entire table as summing to one $\sum_i \sum_j \pi_{ij} = 1$
    - b) Could fix the total # observations
    - c) or let total be an r.v.
  - a) is binomial sample, b) is multinomial sampling, c) is Poisson sampling

Contingency tables

- We focus on tests of hypotheses, it turns out that a similar procedure works for all of the examples
- Null hypothesis H_o specifies a null model (e.g., same proportions in each row)
- Expected counts:
  - Compute expected count for each cell of table under the null model (call this E_{ij})
  - $E_{ij} =$ (row i total)(col j total)/(table total)
  - Why?
    - Consider row = pop, col=outcome.
    - col j total / table total = proportion of outcome j
    - under H0, all pop have same prop., so # with outcome j in pop i = (row i total)(prop. j)

Chi-squared tests

- Chi-squared test statistic compares observed and expected counts across table

$$C = \sum_{\text{cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \sum_{\text{cells}} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

- under H_o, $C \sim \chi^2$ with $(r-1)(c-1)$ degrees of freedom, if sample size is sufficiently large
- Cochran's rule: all $E_{ij} > 1$ and 80% of $E_{ij} > 5$
- traditional rule (all $E_{ij} > 5$) is conservative
- If sample size too small:
    - can combine rows or columns
    - use exact = randomization inference

2×2 table

- The two proportion problem in a 2×2 table

popul	success	failure	total
1	Y_1	$n_1 - Y_1$	n_1
2	Y_2	$n_2 - Y_2$	n_2
total	$Y_1 + Y_2$	$N - Y_1 - Y_2$	$N = n_1 + n_2$

- Expected counts (use first cell as example)

$$E_{11} = n_1(Y_1 + Y_2)/N = n_1 p$$

  where p is pooled sample proportion
- Chi-squared statistic (d.f. = (2-1)(2-1) = 1) is the square of the z statistic comparing the two proportions

χ^2 test and logistic regression

- Notice there is no distinction between "dependent" and "independent" variables in a contingency table.
  - Can interchange rows and columns without changing meaning/interpretation of the table.
  - Logistic regression has a clear distinction between Y and X.
- Consider 2 x 2 table on previous slide
  Comparing P[success — population]
- Logistic regression model, i indicates row. n_i is the row total

$$Y_i \sim \text{Bin}(n_i, \pi_i), \qquad \text{logit } \pi_i = \tau + \gamma_i$$

- Contingency table model, ij indicates cell of the table, C_{ij} is the count in cell ij

$$C_{ij} \sim \text{Poisson}(\lambda_{ij}), \qquad \log \lambda_{ij} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij}$$

χ^2 test and logistic regression

- CT model above fits the counts perfectly. χ^2 is a measure of lack-of-fit of the additive model (all $\alpha\beta_{ij} = 0$).
- γ_i in the LR model and $\alpha\beta_{ij}$ in the CT model parameterize the same quantity: the difference in probabilities between the rows
- Test statistics and p-values usually slightly diff. because LR uses deviance, CT uses χ^2.

$r \times c$ table

- Null hypothesis depends on scenario
- Examples
  - $r \times 2$ table: let $\pi_i =$ prob of success in pop. i ($i = 1, \ldots, r$) and test $H0 : \pi_1 = \pi_2 = \cdots = \pi_r$
  - $2 \times c$ table: let $\{\pi_{ij}, j = 1, \ldots, c\}$ represent the distn of outcomes in popul i ($i = 1, 2$) and test $H0 : \pi_{1j} = \pi_{2j}$ for all j
  - $r \times c$ table: let π_{ij} represent proportion of popul classified into row i, col j and test $H0$: row and col classifications are indep ($H_o : \pi_{ij} = \pi_{i+}\pi_{+j}$)
- Expected counts and d.f. are computed the same way in each case

$r \times c$ table

- Note: chi-squared test may have many d.f., small P values reject H_o but don't tell how it fails
  - can look at chi-squared residuals: (observed - expected)/$\sqrt{\text{expected}}$
  - Same as the Pearson χ^2 residual in a logistic regression
  - or test more focused hypotheses, i.e.
    - compute χ^2 for a subset of rows and columns
    - or, combine rows and / or columns
    - both, analogous to contrasts

$r \times c$ table

- Example:

Counts			Residuals		
50	20	10	1.831	-1.444	-1.021
10	20	10	-2.119	1.671	1.182
10	10	5	-0.596	0.470	0.332

- Overall χ^2 = 15.85, 4 df, p = 0.0032
- Residuals pick out [2,1] entry as unusually low, [1,1] unusually high
- χ^2 on just 2'nd and 3'rd columns: χ^2 = 0, 2 df, p = 1.00

“Continuity correction”

- A detail to be aware of
- Sometimes, test statistic computed as

$$C = \sum_{ij} \frac{(\mid O_{ij} - E_{ij} \mid -0.5)^2}{E_{ij}}$$

- the -0.5 is called a continuity correction
- Motivation:
  - C is computed from integers, so support is a discrete set of values
  - theoretical distribution is continuous (χ^2)
  - -0.5 improves correspondence between the two distributions
- some use always. I never do.
  - reduces power.
  - Effect on α level of test small unless sample sizes are small
  - when you should be doing a randomization test anyway.

goodness-of-fit test

- Sometimes we have a $1 \times c$ table listing counts of different categorical outcomes and wish to compare the observed dn. to a model (e.g., Poisson, Binomial)
- Chi-squared test
  - same test statistic (sum of (obs - exp)^2/exp)
  - expected counts now computed using the hypothesized model
  - degrees of freedom = $c - 1$
  - assumes model completely specified.
    - Does not account for estimating parameters (e.g. $\hat{\lambda}$ in Poisson).
    - Theory exists (Kendall and Stewart, Adv. Theory of Statistics, 4th ed., section 30.11 et seq.)
    - fewer d.f. how many fewer depends on how parameters estimated
    - I don't know any program that computes this.

Fisher's exact test

- Previous methods ALL assume large samples
- Fisher's exact test for comparing two proportions examines n_{11} and computes the exact probability of observing a table as or more extreme assuming the row and column totals stay fixed
- Why fix row and col totals?
  - (they were fixed by design in Fisher's example)
  - but very rare in practice.
  - theory: row and col totals are ancilliary for inferences about odds ratio(s)
  - so condition on observed total even if not fixed
- Hypergeometric distn is the relevant reference distn

$$\Pr(N_{11} = n_{11}) = \frac{n_{1+}!n_{2+}!n_{+1}!n_{+2}!}{N!n_{11}!n_{12}!n_{21}!n_{22}!}$$

Fisher's exact test

- *P*-value is sum of probability for tables as or more extreme, e.g., if we observe

pop	succ.	fail.
1	1	7
2	4	4

then as or more extreme tables are

pop	succ.	fail.
1	1	7
2	4	4

pop	succ.	fail.
1	0	8
2	5	3

- problem for randomization-based inference because set of possible outcomes larger if do not condition on row and col totals. active discussion, no consensus
- traditional solution is to condition and use Fisher's exact test, in spite of possible problems.

Paired data

- What if the data are repeated measurements (e.g., success/failure at time 1 and success/failure at time 2)
- Still get a 2 $\times$ 2 table but now we don't have independent proportions
- Pairing often ignored - bad analysis!
- Correct analysis: A new 2 x 2 table.
  - cross-classify each pair
  - Row = response at time 1,
  - Col = response at time 2
- Notation: let $\pi_{ij} =$ proportion with resp *i* at time 1 and *j* at time 2; take the table total to be *n*

	Time 2		
Time 1	1(+)	2(-)	total
1(+)	π_{11}	π_{12}	π_{1+}
2(-)	π_{21}	π_{22}	
total	π_{+1}		

Paired data

- Diagonals have no information about change over time
- Tests only use pairs with discordant responses: (-,+) or (+,-)
- Under H0: $\pi_1 = \pi_2$, $p_{-+} = p_{+-}$
- Two approaches:
  - Standard (large sample) approach gives
    - $100(1-\alpha)\%$ CI for $\pi_{1+} - \pi_{+1}$ as:

$$p_{1+} - p_{+1} \pm z_{1-\alpha/2} \sqrt{\frac{1}{n}(p_{1+}(1-p_{1+}) + p_{+1}(1-p_{+1}) - 2(p_{11}p_{22} - p_{12}p_{21}))}$$

    - and normally distributed test statistic $z = (n_{12} - n_{21})/\sqrt{n_{12} + n_{21}}$
  - Small sample test of $H_o : \pi_{1+} = \pi_{+1}$ looks only at off-diagonals: use $n_{12} \sim \text{Bin}(n_{12} + n_{21}, 0.5)$ to find P-value (known as McNemar's test - agrees with prev test in large samples)

More complicated models

- What about more complicated models? i.e. binary response with:
  - factorial treatment structure: below
  - continuous X's (regression): logistic regresion
  - random effects: generalized linear mixed models
  - ordered categories: e.g. Yes, somewhat, No. Hard
- A 6 x 2 x 2 table: Seed germination study

		Amount of water					
Cover	Germinate?	1	2	3	4	5	6
No	Yes	22	41	66	82	79	0
	No	78	59	34	18	21	100
Yes	Yes	45	65	81	55	31	10
	No	55	35	19	45	69	90

3 way tables

- Can consider as a logistic regression with 2 factors
- or a 3 way contingency table

$$C_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk}$$

- i indexes cover, j indexes water, k indexes germinated (Y row or N row)
- Don't care about μ, α_i, β_j, γ_k, and $\alpha\beta_{ij}$
- They depend on total # germinated, total # in each column, total # in each row, and # in each cover/water category
- Interactions with γ_k parameterize differences in germination
  - $\alpha\gamma_{ik}$: between cover levels summing over water
  - $\beta\gamma_{jk}$: between water amounts summing over cover
  - $\alpha\beta\gamma_{ijk}$: 2 way interaction between cover and water

3 way tables

- Why even think about such analysis???
  1. It's complicated,
  2. indirect, (interactions with γ)
  3. and ignores the obvious response: germination
- LR can have numerical difficulties fitting factor models
- Will happen for these data because of the 0's at water amount 6.
- LR β for water amount 6 is $-\infty$
- CT analysis avoids such problems.