Importance of the assumptions / consequences of violating them: (reminder)
Independence: crucial / wrong se so wrong p-value, wrong ci.
Equal variance: depends on equality of sample sizes / when unequal, wrong se
Normality: low when same shape. Outliers always a concern

A: Overview of last 5 weeks, how to choose a method
Expanded version of the book's process in section 3.4.
Consistent with the book's recommendations where they overlap
Goal is to provide a justifiable analysis.
No single right way to choose. My suggestion based on:
Question $\rightarrow$ (Design) $\rightarrow$ Data $\rightarrow$ Analysis (makes assumptions)
which we saw in the first week

My suggested approach:
What is the question?
Differences in location (mean, median)? Differences in spread?
What is the study design?
experimental or observational? (i.e., causal conclusions?)
Are data paired or not?
Any concerns about independence: eu = ou? serial effects?
Do I want a confidence interval? or just a p-value?
What assumptions are reasonable (or not badly violated)?
skewed distribution of errors? apparent outliers?
reasonably equal variances?
If you want a ci, use t-based methods,
transform responses to improve assumptions
If assumptions are good, perhaps after transformation
use a t-test
If assumptions are not good even after transformation
use a non-parametric test (on ranks) or randomization test (on data values)
If some observations are censored, or want resistance to outliers
use a non-parametric test (on ranks)

This is the end of the material covered on midterm I.

Seriously non-normal data:
Response is Yes/No (Bernoulli data) or a count (0, 1, 2, $\cdots$)
Discrete responses: 1/0 for Yes/No, integers for counts
If statistics is a collection of named methods, need lots of new names
General principles are identical to what we've already seen (or will see)
details are different
computing much harder, but that's what computers take care of

B: Equality of two proportions:
example Vit C study (Case study 18.2), notation:

Treatment	# not	# cold	Row total
Placebo	76	335	$R_1 = 411$
Vit. C	105	302	$R_2 = 407$
Col. total	$C_1 = 181$	$C_2 = 637$	$N = 818$

Bernoulli data: Response is Yes or No
Focus (usually) on proportion of Yes (or No) within a group
Proportion = # Yes / # tries
Common to code $Y_i = 1$ (Yes) or 0 (No)
Then proportion is the average Y_i, $p = \Sigma Y_i / N$
Percent = 100 $\times$ proportion
Precision: depends on population proportion, π:
se $p = \sqrt{\frac{\pi(1-\pi)}{N}}$
Not constant! (big difference from normally distrib. data)
largest when $\pi = 0.5$ (see figure on board)
estimate of se $p = \sqrt{\frac{p(1-p)}{N}}$ (plug-in p for π)
Inference:
ci for π: $p \pm z_{1-\alpha/2}$ se p
se computed using p, i.e., plugging p into se formula
95% ci: $z_{0.975} = 1.96$
Endpoints can be < 0 or > 1.
Lots of other ways to compute CI for a proportion
One group, test $\pi = \pi_0$: $Z = (p - \pi_0)/\text{se } p$
se computed using π_0, i.e., plugging π_0 into se formula
both use z scores, not t scores, because not estimating s
Z has a normal distribution with mean 0, variance 1
equivalent to T distribution with ∞ d.f.

Bernoulli and Binomial distributions:
Two different ways of describing yes/no data
1) Focus on individuals: $Y = 1$ or 0 i.e. event happened (1) or it didn't (0)
This is a Bernoulli distribution.

Has 1 parameter, the probability of the event, π

$$Y \sim \text{Bernoulli}(\pi)$$

2) Focus on number of times an event "happens" out of N "tries"
This is a Binomial distribution

$$Z \sim \text{Binomial}(N,\ \pi)$$

If N individuals have the same π, number of events is:

$$Z = \sum_{i=1}^{N} Y_i \sim \text{Binomial } (N, \pi)$$

Tests of whether two groups have same proportion, i.e., $\pi_1 = \pi_2$:
Could construct a Z test of $\pi_1 - \pi_2 = 0$
Need to compute se $\hat{\pi}_1 - \hat{\pi}_2$ when Ho true, i.e., $\pi_1 = \pi_2$
Requires P[cold] ignoring treatment group, use total # colds, total # individuals
In terms of above table, overall P[cold] $= \hat{\pi} = C_2/N$

Chi-square test of equal proportions
Chi-square test uses model comparison, instead of a Z test for one parameter
Simpler way to do the computations
Generalizes to more than 2 groups (or more than 2 responses)

C: Model comparison, using T-test as example:
Model I: two groups have the same population mean

$$\begin{array}{ll} \text{Group A} & Y_{Ai} = \mu + \varepsilon_{Ai} \\ \text{Group B} & Y_{Bi} = \mu + \varepsilon_{Bi} \end{array}$$

Model II: two groups have different population means

$$\begin{array}{ll} \text{Group A} & Y_{Ai} = \mu_A + \varepsilon_{Ai} \\ \text{Group B} & Y_{Bi} = \mu_B + \varepsilon_{Bi} \end{array}$$

Model I expresses the null hypothesis of the test
Model II: expresses "not the null hypothesis"
Model II is more flexible, will always fit as well or better
never worse than model I
If Ho is true, model II will fit a little bit better than Model I
If Ho is false (means not the same)
model II will fit a lot better than model I
For normally distributed data, use Sums-of-squared errors as measure of fit
Leads to an F test
Will see all the details when we cover ANOVA and F tests

Model comparison, for yes/no responses:

Model I: two groups have the same proportion of Yes (= had a cold)

$$\begin{array}{rl} \text{Vit C} & Y_{1i} \sim \text{Bernoulli}(\pi) \\ \text{Control} & Y_{2i} \sim \text{Bernoulli}(\pi) \end{array}$$

Model II: two groups have different proportions of Yes

$$\begin{array}{rl} \text{Vit C} & Y_{1i} \sim \text{Bernoulli}(\pi_1) \\ \text{Control} & Y_{2i} \sim \text{Bernoulli}(\pi_2) \end{array}$$

Or:

Model I: two groups have the same proportion of Yes (= had a cold)

$$\begin{array}{rl} \text{Vit C} & Z_1 \sim \text{Binomial}(N_1,\ \pi) \\ \text{Control} & Z_2 \sim \text{Binomial}(N_2,\ \pi) \end{array}$$

Model II: two groups have different proportions of Yes

$$\begin{array}{rl} \text{Vit C} & Z_1 \sim \text{Binomial}(N_1,\ \pi_1) \\ \text{Control} & Z_2 \sim \text{Binomial}(N_2,\ \pi_2) \end{array}$$

Use Chi-square statistic as measure of fit

Because Bernoulli or Binomial data have different properties than Normal data

se $\hat{\pi}$ not constant, not dependent on a separately estimated s

D: Chi-square statistics:

Compare observed counts to what is expected given a model

Observed counts and notation

Treatment	# not	# cold	Row total
Placebo	O_{11}	O_{12}	R_1
Vit. C	O_{21}	O_{22}	R_2
Col. total	C_1	C_2	N

Expected cell counts when Ho true ($\pi = \pi_1 = \pi_2$)

Treatment	# not	# cold	Row total
Placebo	$E_{11} = R_1(1-\pi)$	$E_{12} = R_1\pi$	R_1
Vit. C	$E_{21} = R_2(1-\pi)$	$E_{22} = R_2\pi$	R_2
Col. total	C_1	C_2	N

Remember when Ho is true, estimate $\hat{\pi} = C_2/N$, the overall proportion of the Cold event

Treatment	# not	# cold	Row total
Placebo	$E_{11} = R_1C_1/N$	$E_{12} = R_1C_2/N$	R_1
Vit. C	$E_{21} = R_2C_1/N$	$E_{22} = R_2C_2/N$	R_2
Col. total	C_1	C_2	N

Logic: Ho is $\pi_1 = \pi_2$.

Reject when observed counts (O_{ij}) are far from their expected counts (E_{ij})

Use Chi-square statistic as a measure of fit

$$C = \sum_{ij} \left[\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right]$$

Similar to sum of squares; denominator accounts for unequal variance

Model comparison:

Full model: Two P[cold], one for placebo, one for Vit. C, Fits perfectly, $C = 0$

Reduced model: One P[cold] = π, C computed as above

Large values $\Rightarrow$ observed far from expected, reject Ho

Theory: when $\pi_1 = \pi_2$ and sample size sufficiently large,

$C \sim \chi^2_k$ Chi-square distribution with k df

df = (# Rows - 1) (# Cols -1)

When is sample size sufficiently large?

Common advice: all $E_{ij} \geq 1$ and most (80% +) ≥ 5

When sample size not large, usual small sample procedure is Fisher's exact test

Optional: demonstration that $C = 0$ for the full model (two P[cold])

Will show that under the full model $E_{ij} = O_{ij}$ for every cell

Under the full model, P[cold — group] = proportion of colds in a group

O_{12}/R_1 for placebo group

O_{22}/R_2 for Vit C group

Substituting into table of expected counts:

Treatment	# not	# cold	Row total
Placebo	$E_{11} = R_1O_{11}/R_1$	$E_{12} = R_1O_{12}/R_1$	R_1
Vit. C	$E_{21} = R_2O_{21}/R_2$	$E_{22} = R_2O_{22}/R_2$	R_2
Col. total	C_1	C_2	N

Treatment	# not	# cold	Row total
Placebo	$E_{11} = O_{11}$	$E_{12} = O_{12}$	R_1
Vit. C	$E_{21} = O_{21}$	$E_{22} = O_{22}$	R_2
Col. total	C_1	C_2	N

E: Sampling models: How to obtain the data in above table; three common ways
Prospective Binomial sample: e.g., Vit C data
Two (or more) groups, then observe events, estimate P[event | group]
Retrospective Binomial sample: e.g., case-control study (Case study 18.3).
Especially useful when event is rare
Sample C_1 events and C_2, observe group for each
can estimate P[group | event] but not P[event | group] (without more info)
can estimate odds ratio (see below)
Multinomial sample: e.g. genetic linkage study
observe 4 groups defined by row and column labels
Q concerns independence of the row and column classifications
These differ by what is fixed by the design
Prospective Binomial: Number in each group (row totals)
Retrospective Binomial: Number of events and non-events (column totals)
Multinomial: Total number of subjects (only N)
There are still more sampling models, but they are much less frequently used
Theory $\Rightarrow$ use Chi-square test for all three (when sample size large)
Different small sample methods

F: Odds ratios to describe differences in two proportions:
Difference, $p_1 - p_2$, has issues when applied to a wide range of populations
Vit. C: P[cold | Placebo] = 0.82, P[cold | Vit. C] = 0.74, Estimated diff. is 8%
What if a year or place when colds infrequent, e.g. 6% on placebo.
Would estimate P[cold | Vit.C] in that place as 6% - 8% = -2% ???
Odds ratios quantify relationship between two proportions
that is applicable across a wide range of baseline proportions
Odds = $\pi/(1-\pi)$. related to betting: horse is a 2:1 favorite.
range from 0 (Prob = 0) to ∞ (Prob = 1)
Odds = 1 $\Rightarrow$ Prob = 0.5
Statistical analysis commonly uses log odds
range from $-\infty$ (Prob = 0) to ∞ (Prob = 1)
log Odds = 0 $\Rightarrow$ Prob = 0.5
Odds ratio: comparison between two groups

$$\text{Odds}_1/\text{Odds}_2 = \frac{\pi_1(1-\pi_2)}{(1-\pi_1)\pi 2}$$

Odds ratio = 1 when proportions equal, > 1 when $\pi_1 > \pi_2$
commonly use log odds ratio: = 0 when $\text{Odds}_1 = \text{Odds}_2$ or $\pi_1 = \pi_2$

G: Inference for log odds ratio

$$\text{estimate } = \log\left[\frac{O_{12}\, O_{21}}{O_{11}\, O_{22}}\right]$$

Vit C study:

Choose to use log odds of cold in placebo - log odds of cold in VitC

odds of cold in placebo = $O_{12}/O_{11} = 335/76 = 4.41$

odds of cold in Vit C = $O_{22}/O_{21} = 2.88$

log odds ratio = $\log(4.41/2.88) = \log 1.53 = 0.43$

Easy to misinterpret as odds of “not cold’ in placebo vs in vitC

or odds of cold in Vit C vs in placebo

Sign difference (+ or -1). If it matters, I check against proportions

se $\approx \sqrt{\frac{1}{O_{11}} + \frac{1}{O_{12}} + \frac{1}{O_{21}} + \frac{1}{O_{22}}} = \sqrt{0}.029 = 0.17$

ci for log odds ratio: estimate $\pm z_{1-\alpha/2}$ se

95% ci: $z_{0.975} = 1.96$

exponentiate to get ci for odds ratio

On log odds scale: $0.43 \pm (1.96)(0.17) = (0.096,\ 0.76)$

For odds ratio: $(\exp(0.096),\ \exp(0.76)) = (1.10,\ 2.14)$