A: Assumptions of a randomization test:
independence: errors are independent
Two ways to "get" independence:
observations randomly assigned to treatments, one obs. at at time
i.e., not assigned in groups of observations

Assumptions of a 2 sample t-test:
errors are independent
errors have equal variances
errors are normally distributed

Assumptions of a Wilcoxon rank sum test:
errors are independent
equivalent of equal variance and normality can be stated in various ways
because Wilcoxon has more than one theoretical setup.
Book: same shape and scale. more restrictive than it needs to be
most useful: same shape and scale on some transformed scale.
so errors, appropriately defined, have equal variance but not necessarily normal

B: Residuals and errors:
Assumptions are about errors (deviation from pop. mean), $\varepsilon_{ij} = Y_{ij} - \mu_i$
Estimate errors by the residuals: $r_{ij} = \hat{\varepsilon}_{ij} = Y_{ij} - \overline{Y}_{i.}$
Note: residuals are calculated using group-specific means or averages

C: Assumptions of a paired t-test
Pairs are independent $\Leftrightarrow$ Differences are independent
Observations within each pair are correlated, a good thing not a problem
Differences are normally distributed
No assumption about equal variances
Think why equal variances irrelevant for paired data

Assumption of a Wilcoxon signed rank test or sign test
Pairs are independent $\Leftrightarrow$ Differences are independent

D: Evaluating whether assumptions are reasonable
Can be done informally / graphically
or (sometimes) by formal inference (usually tests)
But logic of a test is backwards
Null hypothesis is that assumption appropriate (e.g. errors are normal)
so rejecting null $\rightarrow$ problem with assumption
but accepting null only $\rightarrow$ no evidence of a problem
does not prove that assumption is appropriate
sample size may be too small to evaluate any choice of error distribution

E: Normality:
Many different tests: generally don't use (prefer informal assessment)
graphical assessment: quantile-quantile (QQ) plot
compare sorted residuals to expected values for a normal population
straight line is good
See plots
remember: on residuals, not observations

F: Equal variance: Formal tests
useful when research question is about variability,
many different tests of equal variance
Most commonly used ones depend critically on restrictive assumptions
When errors are normally distributed,
best test is Folded F or the closely related Bartlett's test
both badly wrong when errors are not normal
Currently preferred methods
Levene's test: t-test of $Z_{ij} =\mid Y_{ij} - \overline{Y}_i \mid$
Brown-Forsythe: t-test of $Z_{ij} =\mid Y_{ij} - \text{median}(Y_i) \mid$
Why these evaluate spread:
Mean of Z measures spread in Y
Z values violate normality and equal variance assumptions
But test results are much more robust than Folded F / Bartlett results

G: Equal variance: informal graphical evaluation
Ratio of sd's or variances
Compute r_s = larger sd / smaller sd or r_v = larger variance / smaller variance
$r_v = (r_s)^2$
Note that $r_s \geq 1$ and $r_v \geq 1$
$r_s \leq 2$, $r_v \leq 4$: equal variance is a reasonable assumption
$r_s \geq 3.1$, $r_v \geq 10$: equal variance is questionable
Residual plot
Y = residual. X = Predicted value, one point for each observation
For t-test, predicted values are the group means: for Y_{ij} that is $\overline{Y}_{i.}$
Looking for approximately same visual spread (vertically) in both groups
Will use residuals plots a lot
Much more useful for more than 2 groups or for regressions

H: Independence:
Concept: error for one observation has no information about errors for any other obs.
Diagnosed by looking at study design
Can use data to evaluate specific types of dependence
no general data-based diagnosis of independence
Common ways to violate independence: serial dependence, cluster dependence
serial dependence: common in data collected over time.
If yesterday higher than predicted, today likely also.

Use time series methods to account for the correlation over time.
cluster dependence: data collected in clusters
example: treatment randomized to school, data collected on students
treatment (to a school) randomized to groups of obs. (students at that school)
common issue in many randomized experiments
Diagnosis: compare experimental and observational units
Experimental unit:"thing" randomized to treatments
Observational unit: "thing" represented by one row of data.
Independence?: Is the o.u. the same as the e.u.?
problem with non-independence when o.u. different from the e.u.
Things to be careful about:
Observational unit is not the response variable
Experimental unit is not the treatment

Example 1:
Randomized experiment, comparing 2 nutrition supplements in cows
Treatments randomly assigned to pens, 10 pens total, 5 per treatment
Each pen has 3 cows
response is average daily weight gain for each cow
Summary: 2 treatments, 10 pens, 30 cows
What is the eu?
what is the ou?
Is there an issue with independence?

Example 2:
Similar study to example 1, except only one cow per pen
What is the eu?
what is the ou?
Is there an issue with independence?

Fixes for non-independence:
5870 solution: average obs. within clusters $\Rightarrow$ one row of data per e.u.
more general solution: model with two sources of variability (e.g. school and student)
called a mixed model, many types, common ones covered in 5710

Example 3:
same study as example 1.
Average 3 cows $\rightarrow$ one average weight gain per pen
What is the eu?
What is the ou?

I: Importance of the assumptions / consequences of violating them:
Independence: crucial.
non-independence → wrong se so wrong p-value, wrong ci.
Equal variance: depends on equality of sample sizes
when unequal → wrong se and/or wrong df
Book's figure
Much less of an issue when equal sample sizes
Normality: low when same shape

J: Treatment of outliers (Display 3.6):
Don't just delete! Hamburger study: 2.5 cfu/gm is the most important value
Book (and my) recommendations:
Is there any error (in measurement, transcription, ...): fix
Is that obs from a different population:
focus on the majority, remove all obs from that minor population
Analyze data with and without the "outliers":
similar results, report with all obs
different results, report both, probably emphasize one set.
CONSORT diagram for flow of individuals through study (see figure)
Various scientific fraud cases where "inconvenient" observations were deleted
but that was hidden when writing the paper

K: Transformations:
apply a function to the response values, e.g., Z = log(Y)
Analyze Z instead of Y
intent is that transformed values, the Z's better fit assumptions
many transformations, many opinions, many advanced alternatives
log(y) is often useful when:
errors are skewed (long upper tail) and/or
group with larger mean has larger variance
NOTE: log(y) is **natural log** = ln(y) on some calculators. NOT log10(y)
backtransformation is exp(z). If $z = \log(y)$, then $y = e^z = \exp(z)$
Note: If you do use log base 10, backtransformation = 10^z

L: Interpretation after log transformation: $Z = \log Y$
average: on log scale, exp(average Z) is geometric mean of observations, Y
median: on log scale, exp(median Z) is median(Y)
Commonly, Z is symmetrical, so median Z ≈ mean Z ⇔ exp(mean Z) is median(Y)
so t-test H0 is equal geometric means or equal medians
difference of average logs: exp(diff) is "multiplicative effect"
se difficult. can't just exponentiate se or sd
CI of diff: exp(CI) is CI for multiplicative effect

Example: Hamburger data

Relevant statistics:

		CFU		log CFU	
Group	n	mean	sd	mean	sd
active	6	0.107	0.072	−2.53	0.948
control	6	0.723	0.885	−0.77	0.953

Results on the log scale
- difference in means (control - active): 1.761
- Pooled sd: 0.950, 10 df
- se for difference: 0.548
- T statistic for H0 diff = 0: 1.761/0.548 = 3.21, p = 0.0093
- ci for difference (control - active): (0.538, 2.984)

Back transformed results
- Multiplicative effect: $\exp 1.761 = 5.818$
- T statistic for H0 ratio = 1: 1.761/0.548 = 3.21, p = 0.0093
- CI for multiplicative effect: $(\exp 0.538,\ \exp 2.984) = (1.71, 19.76)$
  - Notice that the log scale CI is symmetrical around the estimate
  - The backtransformed CI is not, but 5.81/1.72 = 3.4 = 19.9/5.81

What if you took the difference as active - control:
- difference in means, as (active - control): −1.761
- CI for that difference (−2.984, −0.538)

Backtransformed results for active - control:
- multiplicative effect = $\exp -1.761 = 0.17$ Note: 0.17 = 1/5.81
- CI for multiplicative effect = $(\exp -2.984,\ \exp -0.538) = (0.051, 0.58)$

Reporting conclusions about multiplicative effects:

Three ways to word these the estimates:
1) median cfu in control is 5.81 times that the treatment median
2) median cfu in treatment is 0.17 times that the control median
3) median cfu is treatment is 83% less than the control median

Reality check: ratio of data-based medians
- Useful to make sure you know whether effect should be > 1 or < 1
- Medians are control: 0.44 and treatment: 0.11, ratio, as control / active, = 4.0
  - Not same value as that calculated from the log scale difference
  - Because two different ways to estimate the same quantity
  - Correspond to two different models for the data
  - Expect to get different estimates from different models

Inference on the ratio of medians (or the multiplicative effect)

Much easier to do from the log-scale analysis than from the medians of the data

Some things are easy; others are still hard

The hard things:
- standard error of the ratio, i.e., the 5.81 or 4.0 above.
- tests or confidence interval for the data-based ratio (the 4.0 above)

The easy things:

1. Test of ratio = 1 $\Leftrightarrow$ test whether difference of log means = 0
   do a t-test on $\log Y$

2. Confidence interval for the ratio
   Calculate the confidence interval for log difference, exponentiate endpoints

Conclusions about inferences for hamburger bacteria:

1. Test whether median concentration the same in the two treatments:
   t test on $\log Y$. p < 0.0092
   strong evidence that the median concentrations are not equal
   Or, strong evidence that the ratio of bacterial concentrations is not 1

2. Estimate a 95% confidence interval for the ratio of median concentrations
   95% confidence interval for the multiplicative effect is (1.72, 19.9)
   OR: 95% confidence interval for the ratio of medians is (1.72, 19.9)