Final Review

Chapters 1, 2

Controlled experiment: one in which the investigators get to determine who is in the treatment group and who is in the control group. We hope they will choose at random, using chance (“randomized controlled”).

Placebo: A sugar pill or other “fake” treatment that resembles the real treatment but lacks the active ingredient. Used to make the study “blind” so that the subjects do not know which group they are in. This reduces psychosomatic effects and differences in behavior between the groups.

Observational study: one in which the subjects have determined which group they will be in. There may be a "control group" – these are the subjects who did not choose the treatment. Always has confounding factors.

Confounding factor: something that could account for an association and make us doubt that a causal connection exists. Need to explain how the confounding factor relates to *both* factors in the association.

Crosstabs: break down the subjects into smaller groups with similar values of known confounding factors to try to "adjust for" the confounding factor.

Simpson's paradox: patterns for subgroups can differ from overall patterns

Chapter 3

Histograms

- area represents percentages
- total area is 100%
- height = percentage/width
- both axes must have proper scales
- label for y axis is “Percentage per” unit on the x-axis
- area = base x height
- to approximate percentages, assume the data are evenly spread in the intervals and work out areas

Chapter 4

Average, median, r.m.s. size, SD

- what they measure
- how they relate to histograms
- how to calculate them (calculator for AV, SD)
- how they are affected by including more data, removing data, adding a constant, multiplying by a constant, combining two groups, etc.
- longitudinal versus cross-sectional studies

Chapter 5

Normal Curves

- standard units say how many SDs above or below average:

$$z = \frac{x - AV}{SD}$$

- use tables to approximate areas
- percentiles for the normal curve

$$x = \text{average} + z(SD)$$

- in this chapter all methods that use the tables are not
_ valid if the histograms do not follow the normal curve

Chapter 6

Measurement Error

- chance error
- bias
- outliers
- estimate the true value using the average
- the SD gives the likely size of the chance error in a new measurement
- 95% of the time we will be between

  true value – 2(SD)

  true value + 2(SD)

Chapter 7

Lines, Points, Slopes, and Equations

- given two points on a line, $\left(x_1, y_1\right)$ *and* $\left(x_2, y_2\right)$

  slope $= m = \dfrac{y_2 - y_1}{x_2 - x_1}$

- given the slope m and points $\left(x_1, y_1\right)$ *and* $\left(x_2, y_2\right)$

  the equation of the line is

$$y - y_1 = m(x - x_1)$$

Chapter 8

Correlation

- scatter diagrams
- positive, negative, strong, weak
- the correlation coefficient
- approximating the 5 number summary from a plot

$$AV_X\ ,\ SD_X\ ,\ AV_Y\ ,\ SD_Y\ ,\ r$$

Chapter 9

Interpreting Correlation

- r measures ASSOCIATION
- r measures LINEAR association
- r is badly affected by outliers
- ecological correlations are artificially strong
- association is not causation
- correlation is not causation
- the SD line and how to draw it
- calculating y from x if a point is on the SD line

Chapter 10

Regression

- the regression estimate is an estimate of the average of all the y-values for a given x-value
- the regression line and how to draw it
- the six step method to find the regression estimate
- the equation of the regression line
- predicting y from x using regression
- the regression effect and the regression fallacy

Chapter 11

r.m.s. error

- r.m.s. error = $\sqrt{(1 - r^2)}$ (SD_Y)
- for homoscedastic scatter diagrams
- like an SD for the line:
    - 68% of dots are within 1 r.m.s. error of the line
    - 95% of dots are within 2 r.m.s. errors of the line
- also works at a particular value of x

Chapter 12

The Regression Line

- the regression line contains the point (AV_X, AV_Y)

  and has slope $r \times \frac{SD_Y}{SD_X}$

- the equation of the regression line is

$$y - AV_Y = r\frac{SD_Y}{SD_X}(x - AV_X)$$

Chapter 13, 14

Probability

- multiplication rule
- listing the ways, equally likely outcomes
- independence
- multiplication rule for independent events
- mutually exclusive
- addition rule for mutually exclusive events
- the chance something happens at least once is 1 minus the chance it never happens
- probability rules (the world)

Chapter 15

Probability

- binomial coefficient, the number of ways you can choose *k* objects from among *n* objects

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

- repeated trials: suppose we have *n* independent trial, and the probability that event *E* occurs in any given trial is *p* . Then the probability that *E* will occur exactly *k* times is

$$\frac{n!}{k!(n-k)!}\,p^k\,(1-p)^{n-k}$$

-

Chapter 16

The Law of Averages

- as we increase the number of times, the chance error gets bigger and the percentage error gets smaller

- we are more likely to get exactly ZERO chance error (or percentage error) for a few than a lot

Chapter 17

EV for sum of draws and SE for sum of draws

- formulas
- what do they mean?
- normal curve
- 0/1 boxes

Chapter 18

Probability Histograms

- probability histograms represent chances
- the *Normal Approximation*: When drawing a large number of times, at random, with replacement from a box, the probability histogram (in standard units) for the sum of the draws will follow the normal curve.
- the *Normal Approximation* also works for the average of the draws and the percentage of 1’s drawn from a 0-1 box.

Chapter 19

Sample Surveys

- representative sample
- selection bias
- nonresponse bias
- quota sampling
- simple random sample
- cluster sample

Chapter 20

Chance Errors in Sampling

- EV for % , SE for %
- using the normal curve for the sample percentage
- valid if the number of draws is large enough
- actual sample size determines the accuracy, not the proportion of the population we sample (small sample)

Chapter 21

Accuracy of Percentages

- the bootstrap: approximating the box SD using the sample SD
- 95% confidence interval : sample % ± 2 (SE for %)
- valid if the number of draws is large enough

Chapter 23

Accuracy of Averages

- EV for AV of draws, SE for AV of draws
- using the normal curve for the sample average
- 95% confidence interval : sample ave ± 2 (SE for AV)
- valid if the number of draws is large enough

Chapter 26

Tests of significance

- null, alternative
- test statistic, z test
- P-value
- reject or fail to reject
- conclusions
- statistical significance
- two-tailed tests

Chapter 26

Tests of significance

the t test

- used when the number of draws is small and the tickets in the box follow the normal curve
- test statistic: t test
- use SD$^+$
- degrees of freedom = number of draws – 1
- t curve

Chapter 27

Two sample tests

- need 2 independent simple random samples OR

  a randomized, controlled experiment

- null:

- SE for difference (square root law)

- test statistic z =

Chapter 28

The Chi-Square Test

$$\chi^2 = \mathit{sum\ of}\ \frac{(\text{observed frequency - expected frequency })^2}{\text{expected frequency}}$$

- Chi-square test for goodness of fit:
  null hypothesis: chance model holds
  df = number of rows – 1

- Chi-square test for independence of variables:
  null hypothesis: variables are independent, not related
  df = (number of rows – 1) x (number of columns – 1)