1. (a) The *cases* are the residents in the nursing home.

   (b) The factor (herbal tea) is being introduced to the subjects as an action of the researchers. That the level of the explanatory is imposed on cases is the hallmark of an experiment.

   (c) Possible answers:

      Add a control group: a group that isn't receiving herbal tea, for comparison
      Blinding: Cases all receive a hot drink, but are unaware how levels differ
      Randomization: Type of drink received by patient is decided randomly
      Blocking: perhaps on gender, so women are assigned treatments separately from men
      Double-Blinding: Students do not know which patients drink herbal tea.

   (d) B, the visits of the college students seem a likely confounding variable affecting the moods and health of the residents.

   (e) A; it is the *emotinal state* of patients that is being studied.

2. (iii)

3. (a) A possible research question: What portion of ash trees in the midwestern states have been infected?
   For this (my) research question, the sampling method is biased. If the ash borer has infected any trees in the nature preserve, then trees in close proximity (others in the preserve) are more likely to be infected. The trees in the sample are unlikely to be representative of ash trees throughout the midwest, as a result.

   (b) One could put tokens of two different colors, perhaps blue and green, in a bag so that 60% of tokens were green (infected). To simulate *one* sample, you would draw tokens one at a time from the bag with replacement, until 38 draws were made. You would use the count of greens divided by 38 as a single $\hat{p}$. To simjulate the distribution, you would repeat this process many times.

   (c) The rule of thumb for normality is that $np \geq 10$ and $n(1-p) \geq 10$. In this case, $n = 38$ and $p = 0.6$, so

   $$np = (38)(0.6) = 22.8 \qquad \text{and} \qquad n(1-p) = (38)(0.4) = 15.2,$$

   so both are well over 10. We can assume a normal distribution.

   (d) The completed command:

   ```
   x <- do(50000) * rflip(38, prob=0.6)
   ```

   The other two required commands:

   ```
   gf_histogram(~ prop, data=x)
   sd(~ prop, data=x)
   ```

4. (a) A goes with V, B with Z, C with Y, D with W, E with X, and F with U.

   (b) The data is left-skewed, so the *median* is larger.

   (c) E

   (d) B

   (e) Y

   (f) X

5. The mean:

$$\overline{x} \;=\; \frac{1}{4}(8+13+21+34) \;=\; 19.$$

The deviations $(x_j - \overline{x})$ from the mean:

$$8 : 8-19 = -11, \quad 13 : 13-19 = -6, \quad 21 : 21-19 = 2, \quad 34 : 34-19 = 15.$$

Sum of squared deviations from the mean:

$$\sum_{j=1}^{4}(x_j - \overline{x})^2 \;=\; (-11)^2+(-6)^2+2^2+15^2 \;=\; 121+36+4+225 \;=\; 386.$$

Standard deviation:

$$\sqrt{\frac{1}{3}\sum_{j=1}^{4}(x_j - \overline{x})^2} \;=\; \sqrt{\frac{1}{3}\cdot 386} \;=\; 11.343.$$

6. (a) E (b) B (c) J

7.

```
nrow(nutrition)                                     # answer to part (a)
gf_boxplot(Calories ~ Gender, data=nutrition)       # answer to part (b)
gf_point(Cholesterol ~ Calories, data=nutrition)    # answer to part (c)
filter(nutrition, Cholesterol > 200)                # answer to part (d)
lm(Cholesterol ~ Calories, data=filter(nutrition, Cholesterol > 200))
tally(VitaminUse ~ Gender, data=nutrition)          # answer to part (f)
```

8. (a) Histogram
   (b) Side-by-side boxplots or side-by-side histograms
   (c) Scatter plot

9. (a) Yes, there is an association. More specifically, there is a tendency for lower watershed areas to go with lower average *IBI* levels, while higher areas correspond to higher average *IBI*.
   (b) The correlation is *positive*, as the association is a positive one.
   (c) The model is

$$\widehat{\text{IBI}} \;=\; 52.92 + 0.46(\text{area}).$$

   (d) At area 45 km^2,

$$\widehat{\text{IBI}} \;=\; 52.92 + 0.46(45) \;\doteq\; 73.62.$$