Math 58(B) – Jo Hardin
when do I use what??

Below you will find a series of examples. For each example, you should provide the following information:

- What units are being measured (or observed)?
- What variables are being measured? Are they categorical? binary? quantitative?
- Which is the response and which is the explanatory variable (if appropriate)?
- Is it possible that the data were randomly sampled? If no, why not? If yes, what would you then be able to conclude?
- Is it possible that the explanatory variable was randomly assigned? If not, why not? If yes, what would you then be able to conclude?
- Which type of procedure would you follow to address the issue in the example?

(1) Binomial test (or z-test) of a proportion	(2) One sample z-interval for a proportion
(3) One sample t-test of a mean	(4) One sample t-interval for a mean
(5) z-interval for OR / RR	(6) Two sample z-interval for proportions
(7) Two sample t-test of means	(8) Two sample t-interval for means
(9) Two sample randomization test, two binary variables	(10) Two sample z-test of proportions
(11) t-test for linear regression slope	(12) t-interval for linear regression slope
(13) Confidence interval for mean response (linear model)	(14) Chi-square test
(15) Prediction interval for individual observation (linear model)	(16) none of the above are appropriate

- If the choice of procedures is unclear from the information given, state what additional information you would need.
- If you choose a hypothesis test, state the null and alternative hypotheses, and define the relevant population parameter(s).
- If you choose a confidence interval, define the relevant population parameter(s).
- Check any needed technical conditions applicable to the test.

This handout and the following examples were derived from *Workshop Statistics Student Toolkit* by Rossman and Chance

Examples

1. A campus administrator wants to know if some campus groups are more likely to consume alcohol than others. He takes a random sample of 1500 students and classifies them as high risk or low risk drinkers, and whether they belong to a sorority, fraternity, or neither.

2. A researcher wants to determine whether people with "positive attitudes" tend to live longer than those without positive attitudes. He collects data on those who were classified with a positive attitude and those who were not, and he records how long each person lived.

3. A financial investor is interested in whether the number of houses purchased in a particular city is related to the current interest rate. Every day for 3 months, she records the number of houses purchased in the city and also the current interest rate published by the Federal Reserve.

4. A university is trying to determine whether parking is a problem on its campus. The student newspaper contacts a random sample of 200 students and asks whether or not they are frustrated with the parking situation. They want to estimate the proportion of students at the college who are frustrated with the parking situation.

5. A student reporter wants to know whether Democrats were more likely to vote for Candidate NW than Republicans. She takes random samples belonging to the Democratic and Republican parties and asks whether or not they voted for Candidate NW.

6. Doctors are interested in understanding the range of healthy birth weights for full-term babies. A hospital collects data on all the births at the hospital over the year. They measure: birth weight, premature / not, amount of weight mother gained, mother's age, father's age.

7. A new species of lizards is found, and a researcher wants to know their average flight speed. She captures 30 lizards and races them individually on a track in her laboratory.

8. A psychologist wants to know if the amount of time a couple co-habitates before marriage is related to how long the marriage lasts. He selects a random sample of 500 couples and records how many months they lived together before they were married and then the number of years they have been married or were married before divorce.

9. The Best Buy electronics store wants to estimate how much more men tend to spend at the store than women. They collect receipts for a random sample of 100 customers and examine the total amount of the bill.

10. Every year, 900 people in the US die from bicycle accidents. Researchers are interested in understanding the effect of blood alcohol content (BAC) has on the odds of a fatal accidents. For a random sample of bicycle accidents (both fatal and non-fatal) in a given year, whether the BAC is above the legal limit or not is obtained.

Solutions

1. Chi-square test (14)
   observational units = students
   explanatory variable: whether or not the student belongs to a sorority, fraternity, or neither (categorical)
   response variable: whether the student is a high risk or low risk drinker (categorical, binary)
   H_0: no association between Greek membership and alcohol risk level
   H_a: an association between Greek membership and alcohol risk level

   Note: since variable 1 has 3 categories, we can't use a two-sample procedure

   check: make sure that each *expected* cell has at least one observation, and that 80% of the *expected* cells are greater than 5.

2. Two independent samples t-test of means (7) (as long as the people were randomly selected)
   observational units = people
   explanatory variable: whether or not the individual has a positive attitude (categorical, binary)
   response variable: number of years lived (quantitative)
   μ_1 = mean lifetime of those with a positive attitude in the population
   μ_2 = mean lifetime of those without a positive attitude in the population
   $H_0 : \mu_1 = \mu_2$ (no difference in mean lifetime between the two populations)
   $H_a : \mu_1 > \mu_2$ (those with positive attitudes tend to live longer on average)

   check: independent random samples with at least 20 or 30 people each. If sample sizes are small or particularly not normal, the researchers may prefer a randomization test (9).

3. No applicable method (16)
   We can't use regression here because we don't have **independent** observations (the interest rates from day to day depend on each other.) We would use time series techniques to address this problem.

4. One-sample z-interval for a proportion (2)
   observational units = students
   response variable: whether the student considers parking a problem (categorical, binary)
   π = proportion of all students at this college who find parking a problem
   (we want a CI for π)

   check: first check the binomial conditions. Then make sure that $n\pi \geq 10$ and $n(1-\pi) \geq 10$. With a sample size of 200, we would need $\pi < 0.05$ or $\pi > .95$ to violate the conditions. Because this is unlikely, it seems that our conditions will be met.

5. Two sample z-test of proportions (10)
   observational units = students
   explanatory variable: whether the student classifies themselves as Democrat or Republican (categorical, binary)
   response variable: whether or not the student voted for candidate NW categorical, binary)
   π_D = proportion of Democrats at this school who voted for candidate NW
   π_R = proportion of Republicans at this school who voted for candidate NW
   H_0: $\pi_D = \pi_R$ (Rep and Dem were equally likely to vote for candidate NW)
   H_a: $\pi_D > \pi_R$ (Dem were more likely to vote for candidate NW than Rep)

   Note: because we are interested in a one-sided test, the two-sample proportions test is more appropriate than the chi-square test.

check: first check the binomial conditions. Then make sure the sample size is large enough, we need at least 5-10 successes and at least 5-10 failures in each sample.

6. Prediction interval for individual observation (15)
   observational unit - newborn
   response variable: birth weight
   explanatory variables: premature / not, amount of weight mother gained, mother's age, father's age. We first fit the linear model, but the focus is not on estimating or testing the linear model parameters. Instead, we want to be able to build prediction intervals for births at a given set of the explanatory variables.

   check: randomly selected babies (all of them in a given year hopefully representative of births in other hospitals or across other years), would need to see that the residuals are normally distributed with constant variance, and no non-linear patters.

7. One sample t-interval for a mean (4)
   observational units = lizards
   variable (response) - flight speed of the lizard
   μ = true average flight speed for this species of lizard
   (we want a CI for μ)

   check: randomly selected lizards, at least 20-30 or from a normally distributed population

8. t-test for linear regression slope (11)
   observational units = couples
   explanatory variable: length of time the couple co-habitated before marriage (quantitative)
   response variable: number of years the couple's marriage lasted (quantitative)
   $H_0 : \beta_1 = 0$ (no linear assoc btwn length of co-habitation and length of marriage in the pop)
   $H_a : \beta_1 \neq 0$ (a linear assoc btwn length of co-habitation and length of marriage in the pop)

   check: randomly selected observations, LINE conditions

9. Two-sample t-interval for means (8)
   observational units = customers
   explanatory variable: gender (categorical, binary)
   response variable: total amount of bill (quantitative)
   μ_m = average amount spent by men
   μ_w = average amount spent by women
   (we want a CI for $\mu_m - \mu_w$)

   check: randomly selected observations, sample size of at least 30 each

10. z-interval for OR (5)
    observational units = individuals involved in bicycle accidents
    variable 1 (response) - whether their accident was fatal or not (categorical, binary)
    variable 2 (explanatory) - BAC above the legal limit or not (categorical, binary)
    (we find a CI for the OR)

    Our OR represents the multiplicative change in odds of fatalities for a BAC above the legal limit versus a BAC not above the legal limit. If the CI overlaps 1, we will not be able to claim a significant change in odds of fatality for BAC above the legal limit vs. below the legal limit.

    check: no conditions to check except that we have a random sample