

Stat 210 - HW 1 Solutions/Comments (Fall 2009)

1. One-sample methods

- The population parameter of interest is μ , the mean time to complete the degree for students who take a freshman seminar course. The hypothesis that OIR would like to test is $H_o : \mu = 4.7$ vs $H_a : \mu < 4.7$. Remember that we always set tests up so that the null hypothesis is the hypothesis of no difference or no effect. The alternative hypothesis is what we hope to show. (Writing $H_o : \bar{Y} = 4.7$ is completely inappropriate; we know \bar{Y} so we don't need to have a hypothesis about it.)
- The test statistic is $t = (\bar{Y} - 4.7)/(s/\sqrt{n}) = -0.2/(0.6/\sqrt{40}) = -2.11$. The p -value is $P(t_{39} < -2.11) = .02$. (I give precise p -values but you can just give the info from the table in the text, i.e., $.02 < p < .025$ in this case.) The one sentence summary should state the result in terms of the scientific question of interest as in: There is strong evidence that students taking a freshman seminar graduate more quickly than other students. It is not a good summary to say that "we reject H_o because $p < .05$ " because (a) we should not always use $.05$ as a cutoff and (b) it's important to address the scientific question of interest.
- The 95% confidence interval for μ is $\bar{Y} \pm t_{39,.975} s/\sqrt{n} = 4.5 \pm 2.023 * 0.6/\sqrt{40} = (4.31, 4.69)$. Note that the CI does not include the null hypothesis value (4.7) which suggests that we ought to reject the null hypothesis at the $.05$ level (two-sided) (which is $.025$ one-sided). The CI gives the same information about H_o as the test.
- This is an observational study because we did not decide which students took the freshman seminar. These students selected it themselves and may be different (more studious?) than typical students. If we think that they may be more studious (this is a possible confounding variable) then we should re-design the study. Prospectively one could randomly assign some students to freshman seminar and others not and then see how long it takes them to graduate. Retrospectively it is harder to address this confound; we could try to find students with similar backgrounds to the freshman seminar enrollees and use this as a control group.
- With a sample size of $n = 40$ we should be reasonably robust to the kind of skewness we would expect here (values above 10 years are very very rare). Thus I would still believe the results. I apologize if I threw you off by saying "highly-skewed".

2. Two-sample confidence interval

- It does seem appropriate to apply two-sample t procedures in this case. There is no reason to suspect dependence among the samples since there are 89 independent men involved (no evidence of twins, siblings, etc.). The data suggest that the variances in the two populations (diet, exercise) are not very different. We don't know σ_1 and σ_2 but we do have sample estimates which are quite similar. Though we don't have any information about whether the weight losses follow a normal distribution, the sample sizes are large enough to believe that we need not worry about normality.
- First note that $s_p = \sqrt{41 * 3.7^2 + 46 * 3.9^2/87} = 3.81$. (There were many errors here. Errors do happen but you should be able to diagnose a problem if s_p does not appear consistent with the two within group standard deviations. Then the 95% CI for the difference (using the pooled t procedure) is $(7.2 - 5.3) \pm t_{87,.975} s_p \sqrt{1/42 + 1/47} = 1.9 \pm 1.9876 * .808 = (0.3, 3.5)$. We are 95% confident that the average weight loss for dieters exceeds the average weight loss for exercisers by 0.3 to 3.5 kg.
- The only item that changes is the t critical value. For 90% we get $t_{87,.95} = 1.6625$ and the 90% CI is (0.6, 3.2). For 99% we get $t_{87,.995} = 2.6335$ and the 99% CI is (-0.2, 4.0).
- The point of this question is to note that since 0 is outside the 90% interval, the two-sided p -value is $< .10$. It is also outside the 95% interval so we know p -value is $< .05$. But zero is inside the 99% CI which means $p > .01$. Hence we can conclude $.01 < p < .05$ (two-sided).
- As designed the study compares the mean weight loss for men **assigned** to diet and men **assigned** to exercise. These are two populations that we can describe and understand if we work at a weight loss clinic. If we only analyze those that complete the treatment then we are changing the populations to which our results apply. This is a problem here because when a doctor has to make a recommendation he doesn't know whether the person will complete or not. (In the clinical trials world this is known as the distinction between an "intention to treat" and "as actually treated".)

3. Two-sample test - I made one error in describing the problem. The response rate is supposed to be the number of calls per 100 flyers in the parking lot.

- This study is an experiment because treatments are being applied (the color of the flyers being distributed).

- (b) The experimental units are the parking lots. Remember the units are the objects to which the treatment is applied and we are choosing a color for the entire parking lot. The treatment is the color of the flyer. The response is the number of responses per 100 flyers. The population of interest is a bit confusing here – we would like to draw conclusions about individuals (what color they will be more likely to reply to) BUT strictly speaking we really can only draw conclusions about parking lots because that is the unit we worked with. (Of course parking lots are not very interesting, it's only the people that park there that are interesting.)
- (c) Several people mentioned that this would make it a binomial problem and not a normal problem – that's true but that's not really a study design issue. What I was getting at here was that if you did the study in one parking lot, then you only really know about people who park there. If it's a shopping center with expensive stores, then you learn about people who park at such shopping centers (and they may be different than other people).
- (d) $s_p = \sqrt{5 * .6^2 + 5 * .8^2 / 10} = \sqrt{.5} = .71$ and the effect size is $(3.1 - 2.7) / 7.1 = .566$.
- (e) Notice that $t = (\bar{y}_1 - \bar{y}_2) / (s_p \sqrt{1/n_1 + 1/n_2}) = (effect.size) \times \sqrt{n_1 n_2 / (n_1 + n_2)}$.
- (f) The test statistic (using formula from the previous part) is $.566 * \sqrt{3} = 0.98$. This should be compared to a t with 10 d.f.. A two-sided test seems appropriate here because there is no a priori theory as to which color should work better. The p-value is .35 which suggests that there is not much evidence of a color effect.
- (g) The effect size depends only on the means and s.d.s and thus would not change with a bigger sample having the same summary statistics. The test statistic does change because of the second term in (e). If the sample size quadruples then the test statistic would increase (by a factor of the square root of four) and so the p-value would decrease. More precisely here $t = 1.96$ with 46 d.f. which yields a p-value of .06. This is an important point, the same observed difference in sample means becomes more statistically significant (or more convincing evidence) if the sample size is larger.

4. Power calculation I - the normal case

- (a) We reject if $P\left(\frac{\bar{Y}_2 - \bar{Y}_1}{\sigma\sqrt{2/n}} > z_{1-\alpha}\right)$. This means we reject for $\bar{Y}_2 - \bar{Y}_1 > c$ where $c = z_{1-\alpha}\sigma\sqrt{2/n}$.
- (b) Power = $P(\bar{Y}_2 - \bar{Y}_1 > z_{1-\alpha}\sigma\sqrt{2/n})$ where now $\bar{Y}_2 - \bar{Y}_1 \sim N(\delta, 2\sigma^2/n)$;
 Power = $P\left(\frac{\bar{Y}_2 - \bar{Y}_1 - \delta}{\sigma\sqrt{2/n}} > z_{1-\alpha} - \frac{\delta}{\sigma\sqrt{2/n}}\right) = 1 - \Phi\left(z_{1-\alpha} - \frac{\delta}{\sigma\sqrt{2/n}}\right)$.
- (c) No sketch is provided here but note ... when $\delta = 0$ the power is α and then it increases (following the shape of a normal cdf) to one as $\delta \rightarrow \infty$.
- (d) Setting power equal to $1 - \beta$ implies setting $\Phi\left(z_{1-\alpha} - \frac{\delta}{\sigma\sqrt{2/n}}\right) = \beta$ which means that the argument in $\Phi()$ must be equal to $z_\beta = -z_{1-\beta}$. Solving for n yields $n = 2(z_{1-\alpha} + z_{1-\beta})^2\sigma^2/\delta^2$.

5. Power calculation II - categorical data

- (a) If the gender of child can be thought of as a Bernoulli trial with probability 0.5, then the probability of two males equals the probability of zero males equals $0.5 * 0.5 = .25$. The probability of one male and one female child is $2 * 0.5 * 0.5 = .50$ (the factor of two accounts for the fact that the male and female can be born in either order). Thus this makes sense as the null hypothesis.
- (b) There was a typo here which most of you discovered, it should say $p_2 = .25 + \delta/2$. This alternative makes sense (to me) because it preserves the symmetry and of course the probabilities sum to one. Each probability must be between 0 and 1 which means $-0.5 \leq \delta \leq 0.5$. Of course for our theory we know $\delta > 0$.
- (c) I told you this was hard! There are two approaches that I thought of.
- i. Simulation - Please read through carefully because if you understand the idea behind this simulation then you understand power. We first pick an alternative ($\delta = .05$ which means $p_o^* = .275, p_1^* = .45, p_2^* = .275$), a sample size N , and a significance level ($\alpha = .05$ which corresponds to a chi-square statistics cutoff of 5.99). Then we repeat the following many times: (1) generate data representing the number of male children in N households using the alternative probability distribution $p_o^* = p_2^* = .275, p_1^* = .45$; (2) compute the chi-squared test statistic (comparing the simulated data n_o, n_1, n_2 to the proportions $p_o = p_2 = .25, p_1 = .5$ from the null hypothesis); (3) note whether the test would reject H_o or not. After many simulations, the proportion of rejected H_o 's is the power. It turns out see below that you need $N = 1000$ families to have power greater than 0.80 when $\delta = .05$. If $\delta = .10$ (a bigger difference), then you only need 300 families to get power above .80. If $\delta = .20$, then even 100 families is enough to have very high power.
 - ii. Theory - Just to give you some idea about how to think about this The data n_o, n_1, n_2 is random (that's why we simulate). If H_o is true and N is large (doesn't have to be too large here), then we can think of each n_i as a normal random variable with mean Np_i (where p_i is from null distribution) and variance $Np_i(1 - p_i)$.

It's this kind of logic that leads to the chi-square distribution (we are squaring normals and adding them together). If H_a is true and N is large, then we can think of each n_i as a normal random variable with mean Np_i^* (where p_i^* is the alternative distribution) and variance $Np_i^*(1 - p_i^*)$. Let's consider one of the chi-square terms, say the chi-square term that corresponds to n_o . If H_a is true then n_o has mean $N(.25 + \delta/2)$ which means $(n_o - N(.25))$ is approximately normal with mean $\delta/2$. Thus the chi-square contribution from this term is about $(N^2\delta^2/2^2)/N(.25) +$ another term that looks like the chi-square contribution when H_o is true. This is not 100% accurate because it misses a term but it gives you some idea about what is going on. If you do this for the 3 terms in the chi-square statistic, then you end up believing that the test statistic is $4N\delta^2$ PLUS a chi-square random variable with 2 d.f.. Then you need to pick N so that $4N\delta^2$ will shift a chi-square distribution far enough to give the desired power. This approach uses some theory work and gets you a simple probability calculation for the APPROXIMATE power (instead of the simulation). I have shown the results of this approximation in the last column of the table of simulation results below. You will see that this crude theoretical approximatoin doesn't match the simulations very well but it does give an answer in the right ballpark.

SIMULATION RESULTS

```
FIRST -- A SIMULATION PROGRAM THAT IS EASY TO FOLLOW (it has a loop that is executed for each of the mc trials)
chipower <- function(n,delta,mc){
p0 <- c(.25,.5,.25)
p1 <- c(.25 + delta/2, .5 - delta, .25+ delta/2)
nn <- c(0,0,0)
out <- rep(0,mc)
for (i in (1:mc)) {
nn[1] <- rbinom(1,n,p1[1])
nn[2] <- rbinom(1,n-nn[1],p1[2]/(1-p1[1]))
nn[3] <- n - nn[1] - nn[2]
chi <- sum((nn-n*p0)*(nn-n*p0)/(n*p0))
out[i] <- ifelse(chi > qchisq(.95,2),1,0)
}
sum(out)/mc
}
```

```
SECOND -- A MORE EFFICIENT SIMULATION PROGRAM (does all mc trials at once)
chipower2 <- function(n,delta,mc=1000){
p0 <- c(.25,.5,.25)
p1 <- c(.25 + delta/2, .5 - delta, .25+ delta/2)
n0 <- rbinom(mc,n,p1[1])
n1 <- rbinom(mc,n-n0,p1[2]/(1-p1[1]))
n2 <- n - n0 - n1
chi <- (n0 - n*p0[1])*(n0-n*p0[1])/(n*p0[1]) +
(n1-n*p0[2])*(n1-n*p0[2])/(n*p0[2]) +
(n2-n*p0[3])*(n2-n*p0[3])/(n*p0[3])
sum(chi > qchisq(.95,2))/mc
}
```

SIMULATION RESULTS (Based on 10,000 simulated data sets)

```
FOR DELTA = .05
  N delta power approx
  100 0.05 0.1395 0.0801
  200 0.05 0.2433 0.1260
  300 0.05 0.3283 0.2230
  400 0.05 0.4184 0.3702
  500 0.05 0.5110 0.6092
  600 0.05 0.5950 1.0000
  700 0.05 0.6557 1.0000
  800 0.05 0.7257 1.0000
  900 0.05 0.7758 1.0000
  1000 0.05 0.8175 1.0000
FOR DELTA = .10
  100 0.10 0.4242 0.3725
```

```
200 0.10 0.7324 1.0000
300 0.10 0.8919 1.0000
400 0.10 0.9606 1.0000
500 0.10 0.9856 1.0000
FOR DELTA = .20
100 0.20 0.9644 1.0000
200 0.20 0.9999 1.0000
300 0.20 1.0000 1.0000
```