# Homework #1
# Stat 212A, Fall 2015: Topics in Selective Inference
## Instructor: Will Fithian
## Assigned Sep. 17, 2015. Due 11:59pm Oct. 6, 2015

You are welcome to work with each other or consult articles or textbooks online, but you should then go away and write up the problem by yourself. If you collaborate or use other resources, please list your collaborators and cite the resources you used. Please show your work and include code where appropriate.

You can turn in the problem set in class Oct. 6 or under my door (Evans 301) Tuesday night.

**1. Derived Intervals for $\overline{\mu}$** You are working with a scientist, who has one-way layout data:

$$y_i = \mu_i + \varepsilon_i, \quad i = 1, \ldots, m, \quad \text{with} \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0,1),$$

The scientist asks you to come up with FWER-controlling confidence intervals for $\mu_1, \ldots, \mu_m$. You construct

$$C_i = y_i \pm z_{\tilde{\alpha}_m/2}, \quad \text{where} \quad \tilde{\alpha}_m = 1 - (1-\alpha)^{1/m}. \tag{1}$$

(a) After the scientist sees the results, she notices an interesting fact: even though only a few of the intervals exclude 0, most of the $y_i$ are larger than zero. This makes her curious about the value of the parameter

$$\overline{\mu} = \frac{1}{m}\sum_{i=1}^{m} \mu_i,$$

and she expresses regret that she didn't think of asking about it before seeing the data. "Aha!" you exclaim, "but we *can* use the intervals we just constructed to derive an interval for $\overline{\mu}$."

(i) Give an explicit expression akin to (1) for the interval you report.
**Solution:** We can use derived intervals. As long as $\mu_i \in C_i$ for every $i$ (which happens with probability $1-\alpha$), we also have

$$\overline{\mu} \in \inf S, \sup S, \quad \text{where } S = \left\{ \frac{1}{m}\sum_i \mu_i : \ \mu_i \in C_i, \ \forall i \right\}$$

We have

$$\inf S = \frac{1}{m}\sum_i (y_i - z_{\tilde{\alpha}_m/2}) = \overline{y} - z_{\tilde{\alpha}_m/2}$$

and similarly $\sup S = \overline{y} + z_{\tilde{\alpha}_m/2}$, giving interval $\overline{y} \pm z_{\tilde{\alpha}_m/2}$.

(ii) What is the approximate asymptotic radius of this interval as $m \to \infty$?[1]
**Solution:** We showed in class that $z_{\tilde{\alpha}/2}$ is approximately $\sqrt{2\log m}$.

(iii) What is its radius for $\alpha = 0.05$ and $m = 3$, 5, 10, and 100? (Give numbers e.g. 3.45).
**Solution:** See R output below:

```
> m.vals <- c(3, 5, 10, 100)
> alpha.tilde <- 1 - (1-.05)^(1/m.vals)
> qnorm(alpha.tilde/2, lower.tail=FALSE)
[1] 2.39 2.57 2.80 3.47
```

[1] Remember, the radius of the interval is half its width; e.g. the width of the Šidák interval is $2z_{\tilde{\alpha}/2}$ while the radius is $z_{\tilde{\alpha}/2}$.

(b) Your collaborator explains in her paper that she got interested in $\overline{\mu}$ only after looking at the Šidák intervals. One of the referees cries foul: he claims that she is guilty of data dredging and she should remove the interval for $\overline{\mu}$ from the paper. Do you agree that the finding is not properly adjusted for multiplicity? Why or why not?

**Solution:** There is nothing wrong with what your collaborator has done. Derived intervals control the FWER even if we make a post-hoc decision about what contrasts to derive intervals for. The algorithm "1: construct level-$\alpha$ FWER-controlling intervals; 2: derive intervals for any linear combination of $\mu_1, \ldots, \mu_m$ that we want;" controls the FWER at level $\alpha$ no matter how we decide what linear combinations to construct intervals for in step 2. The only way for the derived intervals to be wrong is if one of the intervals in step 1 is wrong, and that happens with probability less than $\alpha$, by construction.

(c) Same question as (a), except suppose that instead of Šidák intervals for all $\mu_i$, the scientist asked you initially to construct Scheffé intervals for all linear contrasts. Then, as in (a), she gets interested after the experiment in $\overline{\mu}$ and wants an interval for it.

(i) Give an explicit expression akin to (1) for the interval you report.
**Solution:** If we made Scheffè intervals, then we have already made an interval for $\overline{\mu}$. Let $1_m \in \mathbb{R}^m$ denote the vector with every coordinate equal to 1, and note that $\overline{\mu} = \nu'\mu$, where $\nu = \frac{1}{m}1_m$. Thus the Scheffé interval for $\overline{\mu}$ is $\overline{y} \pm \|\nu\|\chi_m(\alpha) = \overline{y} \pm \frac{1}{\sqrt{m}}\chi_m(\alpha)$.

(ii) What is the approximate asymptotic radius of this interval as $m \to \infty$?
**Solution:** We showed in class that $\chi_m(\alpha)$ is approximately $\sqrt{m}$, so the interval radius converges to 1.

(iii) What is its radius for $\alpha = 0.05$ and $m = 3$, 5, 10, and 100?
**Solution:** See R output below:

```
> m.vals <- c(3, 5, 10, 100)
> sqrt(qchisq(1-.05, m.vals))/sqrt(m.vals)
[1] 1.61 1.49 1.35 1.12
```

(d) If you were to redo the analysis for a new data set, knowing ahead of time that the scientist is going to be interested in $\overline{\mu}$ as well as the univariate means $\mu_i$, how could you devise a more powerful FWER-controlling procedure than the one you used here? (**Note:** There could be more than one right answer).

(i) Explain how you would construct the intervals $C_i$ for $\mu_i$ and $C_0$ for $\overline{\mu}$, and give a relatively explicit expression for their lengths (e.g. in terms of a quantile of a random variable you can simulate).
**Solution:** We can form simultaneous confidence intervals for the contrasts $N = (\nu_1, \ldots, \nu_{m+1})$ where $\nu_i = e_i$, the $i$th coordinate basis vector, for $i \leq m$, and $\nu_{m+1} = \frac{1}{m}1_m$. Then we can set

$$C_i = \nu_i'y \pm \|\nu_i\| r_\alpha,$$

taking $r_\alpha$ to be the upper $\alpha$ quantile of the random variable

$$R = \max_{i=1}^{m+1} \frac{|\nu_i'\varepsilon|}{\|\nu_i\|},$$

whose distribution we can simulate. Then the interval for $\overline{\mu}$ is

$$C_{m+1} = \nu_{m+1}'y \pm \|\nu_{m+1}\| r_\alpha = \overline{y} \pm \frac{1}{\sqrt{m}} r_\alpha$$

(ii) What is the approximate asymptotic radius of this interval as $m \to \infty$?
**Solution:** The radius should be no smaller than $z_{\tilde{\alpha}_m/2} \approx \sqrt{2\log m}$, since $z_{\tilde{\alpha}_m/2}$ is the upper $\alpha$ quantile of the largest of the first $m$ contrasts. Meanwhile, note that $(\nu_i'\varepsilon)/\|\nu_i\| \sim N(0,1)$ for each

$i$, so the Bonferroni radius of $z_{\alpha/2(m+1)} \approx \sqrt{2\log(m+1)}$ is conservative and must be a little wider than $r_\alpha$.

Because $\sqrt{2\log m} \approx \sqrt{2\log(m+1)}$ for large $m$, then $r_\alpha$ is roughly $\sqrt{2\log m}$, and the radius is roughly $\sqrt{\frac{2\log m}{m}}$.

(iii) What is its radius for $\alpha = 0.05$ and $m = 3$, 5, 10, and 100?

**Solution:** See R output below:

```
> m.vals <- c(3, 5, 10, 100)
> r.alpha <- numeric()
> for(m in m.vals) {
+     n <- 10001
+     sim.y <- matrix(rnorm(n*m),n)
+     sim.R <- apply(sim.y, 1, function(y) max(abs(y), abs(mean(y)*sqrt(m))))
+     r.alpha[as.character(m)] <- quantile(sim.R, probs=.95)
+ }
> r.alpha # Radius for i <= m (||nu_i|| = 1)
   3    5   10  100
2.46 2.62 2.83 3.46
> r.alpha/sqrt(m.vals) # Radius for i = m+1 (||nu_{m+1}|| = 1/sqrt(m))
    3     5    10   100
1.421 1.172 0.895 0.346
```

(e) Same question as (a), except suppose that instead of Šidák intervals for all $\mu_i$, the scientist asked you initially to construct Tukey's HSD intervals for all pairwise comparisons. Show that the derived confidence interval for $\overline{\mu}$ has infinite length.

**Solution:** Let $C = \{\mu : \ \mu_i - \mu_j \in (y_i - y_j) \pm r_\alpha\}$ denote the confidence region of all $\mu$ that are not rejected by any of the $\binom{m}{2}$ Tukey HSD intervals. Note that, for any real number $a$, $\mu^{(a)} = y + a1_m \in C$, because $\mu_i^{(a)} - \mu_j^{(a)} = y_i - y_j$ for every $(i, j)$. But then,

$$\sup_{\mu \in C} \frac{1}{m} \sum \mu_i \geq \sup_{a \in \mathbb{R}} \frac{1}{m} \sum (y_i + a) = \infty,$$

and similarly $\inf_{\mu \in C} \frac{1}{m} \sum \mu_i = -\infty$. Hence the derived interval for $\overline{\mu}$ has infinite radius.

**2. PoSI vs. Scheffé** For regression with $p$ variables and $n$ observations, with known $\sigma^2 = 1$, prove that the PoSI interval radius $r_\alpha$ is always strictly smaller than $\chi_p(\alpha)$ (which is roughly $\sqrt{p}$).

**Solution:** Let $N = \{\nu_{j \cdot M} : \ M \subseteq \{1, \ldots, p\}, j \in M\}$, and recall that $r_\alpha$ is the upper $\alpha$ quantile of the random variable

$$R = \max_{\nu \in N} \frac{|\nu' \varepsilon|}{\|\nu\|}$$

First, we observe that for all $M \subseteq \{1, \ldots, p\}$ and $j \in M$, we have

$$\begin{aligned} \nu_{j \cdot M} &\propto X_{j \cdot M} \\ &= X_j - \mathcal{P}_{X_{M \setminus j}} X_j \\ &\in \text{span}(X). \end{aligned}$$

Set $d = \dim(\text{span}(X)) \leq \min(n, p)$, and let $U = [u_1, \ldots, u_d]$ be a matrix whose columns form an orthonormal basis of span$(X)$. Then $U'\varepsilon \sim N(0, I_d)$, $U'U = I_d$, and for $\nu \in \text{span}(X)$, $UU'\nu = \nu$.

Define $W$ as

$$\begin{aligned} W &= \sup_{\nu \in \text{span}(X)} \frac{|\nu'\varepsilon|}{\|\nu\|} \\ &= \sup_{\nu \in \text{span}(X)} \frac{|\nu' U U'\varepsilon|}{\|U'\nu\|} \\ &= \sup_{\theta \in \mathbb{R}^d} \frac{|\theta' U'\varepsilon|}{\|\theta\|} \\ &= \|U'\varepsilon\| \\ &\sim \chi_d, \end{aligned}$$

Because $R \leq W$, $r_\alpha \leq \chi_d(\alpha) \leq \chi_p(\alpha)$.
Next we show why $r_\alpha < \chi_p(\alpha)$. Let $\Theta = \{U'\nu : \nu \in N\}$, and note that

$$R/W = \max_{\theta \in \Theta} \frac{|\theta' U'\varepsilon|}{\|\theta\| \, \|U'\varepsilon\|} \leq 1,$$

$R/W = 1$ if and only if $\|U'\varepsilon\|$ is exactly proportional to one of the finitely many $\nu \in N$, which occurs with probability 0. Thus, for some $\delta > 0$ we have

$$\mathbb{P}(R/W < 1 - \delta) > \delta$$

Furthermore, $R/W$ is a function of $U'\varepsilon/\|U'\varepsilon\|$, which is independent of $\|U'\varepsilon\|$, a property of the multivariate Gaussian distribution. Thus,

$$\begin{aligned} \mathbb{P}(R \leq \chi_d(\alpha)) &\geq \mathbb{P}(W \leq \chi_d(\alpha)) + \mathbb{P}(W/\chi_d(\alpha) \in (1, (1-\delta)^{-1}],\ R/W \leq 1 - \delta) \\ &= \alpha + \delta \mathbb{P}(W/\chi_d(\alpha) \in (1, (1-\delta)^{-1}]) \end{aligned}$$

which is strictly larger than $\alpha$; hence the upper $\alpha$ quantile of $R$ is strictly smaller than $\chi_d(\alpha)$.

**3. Closing ANOVA** We saw that closing the Simes and Bonferroni procedures resulted in pretty good FWER-controlling multiple-testing procedures (Hochberg's and Holm's procedures, respectively). A natural question to ask is, what if we closed the ANOVA test of the intersection null?

It turns out this is a pretty bad idea! Assume we have the scenario in class where

$$\mu_1 = \cdots = \mu_{k_m} = \rho_m, \quad \mu_{k_m+1} = \cdots = \mu_m = 0$$

(a) Show that even with $O(m)$ non-nulls (quite dense), we need $\rho_m$ to be on the order of $\sqrt{m}$ to get any rejections. More precisely, assume that $k_m = m/2$, and that $\rho_m = o(\sqrt{m})$. Show that $\mathbb{P}(\text{any rejections}) \to 0$ as $m \to \infty$.

**Solution:** The problem is that there are $O(m)$ null observations with relatively small values of $|y_i|$, and any non-null signal will have to generate a $\chi^2$-test rejection even when combined with the small observations.

Let $I_S = \{i : y_i^2 < 1/4\}$. If $\phi_{I_S \cup \{i\}} = 0$ for every $i = 1, \ldots, m$, then the closed-test procedure will not reject any $H_i$.

To get $\phi_{I_S \cup \{i\}} = 1$ we must have

$$\sum_{j \in I_S \cup \{i\}} y_j^2 \geq \chi^2_{|I_S \cup \{i\}|}(\alpha) > |I_S|.$$

In that case, we have

$$y_i^2 > |I_S| - \sum_{j \in I_S} y_j^2 > |I_S| - |I_S|/4$$

So we will not reject the set $I_S \cup \{i\}$ unless $y_i^2 > 3|I_S|/4$. Therefore the whole procedure doesn't reject any $H_i$ unless

$$\max_i y_i^2 > 3|I_S|/4$$

Now, because

$$\mathbb{P}(\varepsilon_i^2 < 1/4) = \Phi(1/2) - \Phi(-1/2) \approx 0.38,$$

and there are $m - k = m/2$ null observations,

$$\begin{aligned} |I_S| &\geq \#\{i > m/2 : \ \varepsilon_i^2 < 1/4\} \\ &\sim \text{Binom}(m/2, 0.38) \end{aligned}$$

So as $m \to \infty$, $\mathbb{P}(|I_S|/m < 1/6) \to 0$.

Next, because $\rho_m < \sqrt{m}/6$ for sufficiently large $m$, we have

$$\mathbb{P}(\max_i |y_i| > \sqrt{m}/3) \leq \mathbb{P}(\max_i |\varepsilon_i| > \sqrt{m}/6) \to 0$$

Tying it all together,

$$\begin{aligned} \mathbb{P}(\text{any rejections}) &\leq \mathbb{P}(\max_i y_i^2 > 3|I_S|/4) \\ &\leq \mathbb{P}(\max_i y_i^2 > m/9) + \mathbb{P}(3|I_S|/4 < m/8) \\ &= \mathbb{P}(\max_i |y_i| > \sqrt{m}/3) + \mathbb{P}(|I_S| < m/6) \qquad \to 0 + 0 \end{aligned}$$

(b) Show that if instead we used Bonferroni's procedure with $k_m = m/2$ and $\rho_m \geq \delta\sqrt{2\log m}$ for any constant $\delta > 0$, then $\mathbb{P}(\text{any rejections}) \to 1$ as $m \to \infty$.

**Solution:** From class, we know the largest of the $m/2$ non-null observations will be roughly of size $\rho_m + \sqrt{2\log(m/2)} \approx (1+\delta)\sqrt{2\log m}$ for sufficiently large $m$. But $z_{\alpha/2m} \approx \sqrt{2\log m}$, so $\max_i |y_i| > z_{\alpha/2m}$ with high probability.

**4. Testing Hypotheses in Fixed Order** Suppose someone gives us an *a priori* ordering on hypotheses $H_{0,1}, \ldots, H_{0,m}$ with $p$-values $p_1, \ldots p_m$ (i.e. the order is specified in advance of looking at the data). We then use the following procedure: If $p_1 \geq \alpha$, stop and accept all null hypotheses. Otherwise, reject $H_{0,1}$ and keep going. Then, if $p_2 \geq \alpha$, stop and accept $H_{0,2}$ through $H_{0,m}$. Otherwise, reject $H_{0,2}$ and keep going, etc. In other words, if $k$ is the index of the first $p$-value that is larger than $\alpha$, we reject $H_{0,i}$ for each $i < k$ and accept the rest.

(a) Prove that this procedure controls the FWER, regardless of the dependence structure of the $p$-values.

**Solution:** Let $i_0^* = \min I_0$, the index of the first true null hypothesis in the list. We cannot make any mistakes unless $p_{i_0^*} \leq \alpha$, which happens with probability at most $\alpha$.

(b) **Challenge** (Optional) Can you formulate this problem as a special case of a closed-test procedure? That is, what intersection-null test is it the closure of? (**Note:** this part suffices to prove part (a) so you can just write "see answer to (b)").

**Solution:** This procedure is a closed-test procedure where we use $\phi_I(p) = 1\{p_{\min I} < \alpha\}$ as our intersection-null test. That is, we reject $H_I$ (in step 1) if $p_{\min I} < \alpha$.

We need to show that

this closed-test procedure rejects $H_i$ $\iff$ the ordered testing procedure above rejects $H_i$

The ordered testing procedure rejects $H_i$ if and only if all of the first $i$ $p$-values are $\leq \alpha$. If $p_j > \alpha$ for some $j \leq i$, then the intersection null test doesn't reject for $\{j, i\}$ because $p_{\min\{j,i\}} = p_j > \alpha$; thus the closed-test procedure does not reject $H_i$. Conversely, if $p_j \leq \alpha$ for all $j \leq i$, then for any $I \ni i$, $\min I \leq i$ and therefore $p_{\min I} \leq \alpha$; thus the closed-test procedure does reject $H_i$.

**5. A Bit of Philosophy** (**Note:** Graded for completion only; write as little or as much as you want, but write something. Also note I don't know the answer to this question!)

Suppose a journal decides to embrace statistical rigor and requires that in each submitted paper, all of the hypothesis tests / confidence intervals, taken together, must control the FWER at level $\alpha = 0.05$. In other words, if ten confidence intervals appear in your paper, they must have been generated according to a procedure guaranteeing that, with 95% probability, all ten cover their true parameters.

In a meeting of the editors, one particularly conservative editor pipes up saying "this is a good start, but really we should be controlling the FWER across all of the inferences in all of the articles in each issue of the journal." Discuss the feasibility of this proposal. Aside from feasibility, do you think this is a good goal? Why stop at FWER for each issue, as opposed to FWER control for each year, or over the entire life of the journal? If you think these proposals are too conservative, is there a principled reason to require FWER control for each article but not for each issue of the journal?

**Solution:** I think arguably there is not really a principled reason to do it for each article but not for each journal issue: just as the "best" results in a study will be singled out to be highlighted in the journal version, it is also true that the "best" findings in the whole journal will be singled out and perhaps reported widely in the media. In that sense, I think that "all $p$-values in the journal" is arguably a relevant family.

In my opinion, this sort of conundrum argues for the "selective" error rates we will study during most of the rest of the class: it is OK if some of the $p$-values in the article, or in the journal are wrong, even if many are wrong. The problem comes when the *highlighted* (selected) findings do not satisfy their advertised frequency properties.