ST520: Statistical Principles of Clinical Trials

HW 3 Solutions

Problem 1

(a) Under the null hypothesis, the assignment of the responses to the treatments is random, so we have $2^4 = 16$ different assignments of responses of outcomes to consider, as there are 2 ways to assign treatment labels to each pair of responses within each block.

These 16 different assignments give rise to 16 different values of $\bar{Y}_B - \bar{Y}_A$, which are tabulated below. Here, we write the treatment labels of the responses in each block and the corresponding value of $\bar{Y}_B - \bar{Y}_A$.

0.3	0.9	1.0	0.2	0.5	0.7	0.1	0.2	$\bar{Y}_B - \bar{Y}_A$
A	B	A	B	A	B	A	B	$\frac{0.1}{4} = 0.025$
						B	A	$\frac{-0.1}{4} = -0.025$
				B	A	A	B	$\frac{-0.3}{4} = -0.075$
						B	A	$\frac{-0.5}{4} = -0.125$
		B	A	A	B	A	B	$\frac{1.7}{4} = 0.425$
						B	A	$\frac{1.5}{4} = 0.375$
				B	A	A	B	$\frac{1.3}{4} = 0.325$**
						B	A	$\frac{1.1}{4} = 0.275$
B	A	A	B	A	B	A	B	$\frac{-1.1}{4} = -0.275$
						B	A	$\frac{-1.3}{4} = -0.325$
				B	A	A	B	$\frac{-1.5}{4} = -0.375$
						B	A	$\frac{-1.7}{4} = -0.425$
		B	A	A	B	A	B	$\frac{0.5}{4} = 0.225$
						B	A	$\frac{0.3}{4} = 0.075$
				B	A	A	B	$\frac{0.1}{4} = 0.025$
						B	A	$\frac{-0.1}{4} = -0.025$

This gives the sampling distribution that $\bar{Y}_B - \bar{Y}_A$, whose pmf we write below:

$$f(\bar{Y}_B - \bar{Y}_A) = \begin{cases} \frac{1}{16} & \text{if } |\bar{Y}_B - \bar{Y}_A| = 0.075, 0.225, 0.275, 0.325, 0.375, \text{or } 0.425 \\ \frac{1}{8} & \text{if } |\bar{Y}_B - \bar{Y}_A| = 0.025 \\ 0 & \text{otherwise} \end{cases}$$

(b) The p-value of our test statistic is $P(\bar{Y}_B - \bar{Y}_A \geq 0.325) = \frac{3}{16} = 0.1875$, as our alternative hypothesis, $H_A : \bar{Y}_B > \bar{Y}_A$, is one-sided.

(c) The main advantage of permuted block randomization is that it usually produces more balanced designs than simple randomization. It is, however, harder to implement than simple randomization, and it is possible for clinicians to bias results by forcing patients into certain treatments based on their order in any particular block, thought this is mitigated by randomizing the size of the block.

Problem 2

(a) To prove that $\left(\frac{n}{n-1}\right) p(1-p)$ is unbiased for $\pi(1-\pi)$, we evaluate its expected value:

$$
\begin{aligned}
E\left[\left(\frac{n}{n-1}\right)p(1-p)\right] &= E\left[\left(\frac{n}{n-1}\right)\frac{X}{n}(1-\frac{X}{n})\right] \\
&= \frac{1}{n(n-1)}E[nX-X^2] \\
&= \frac{1}{n(n-1)}(nE[X]-E[X^2]) \\
&= \frac{1}{n(n-1)}(nE[X]-(E[X^2]-\{E[X]\}^2+\{E[X]\}^2)) \\
&= \frac{1}{n(n-1)}(nE[X]-(Var[X]+\{E[X]\}^2))
\end{aligned}
$$

Here, because $X|n,\pi \sim b(n,\pi)$, we have $E[X]=n\pi$ and $Var[X]=n\pi(1-\pi)$, so substituting these values gives

$$
\begin{aligned}
E\left[\left(\frac{n}{n-1}\right)p(1-p)\right] &= \frac{1}{n(n-1)}(n(n\pi)-(n\pi(1-\pi)+(n\pi)^2)) \\
&= \frac{1}{n(n-1)}(n^2(\pi-\pi^2)-n\pi(1-\pi)) \\
&= \frac{1}{n(n-1)}(n(n-1)\pi(1-\pi)) \\
&= \pi(1-\pi)
\end{aligned}
$$

Thus, $\left(\frac{n}{n-1}\right)p(1-p)$ is unbiased for $\pi(1-\pi)$.

(b) For our hierarchical problem, we know from part (a) that $E[\frac{n_i}{n_i-1}p_i(1-p_i)|n_i,\pi_i]=\pi(1-\pi)$. This holds for all i. Then we can construct a predictor for $E\{\pi(1-\pi)\}$ by averaging the individual predictors for each $\pi(1-\pi)$, namely $\frac{1}{N}\sum_{i=1}^{N}\frac{n_i}{n_i-1}p_i(1-p_i)$.

We have

$$
\begin{aligned}
E[\frac{1}{N}\sum_{i=1}^{N}\frac{n_i}{n_i-1}p_i(1-p_i)] &= \frac{1}{N}\sum_{i=1}^{N}E[\frac{n_i}{n_i-1}p_i(1-p_i)] \\
\text{Use law of iterated expectations} &= \frac{1}{N}\sum_{i=1}^{N}E[E[\frac{n_i}{n_i-1}p_i(1-p_i)|n_i,\pi_i]] \\
&= \frac{1}{N}\sum_{i=1}^{N}E\{\pi_i(1-\pi_i)\} \\
&= \frac{1}{N}NE\{\pi_i(1-\pi_i)\}=E\{\pi_i(1-\pi_i)\}
\end{aligned}
$$

Thus, our estimator $\frac{1}{N}\sum_{i=1}^{N}\frac{n_i}{n_i-1}p_i(1-p_i)$ is unbiased for $E\{\pi_i(1-\pi_i)\}$.

Problem 3

(a) We have, from the law of total variance, $var\{\bar{Y}_i\}=E\{var[\bar{Y}_i|\mu_i,\sigma_i^2,n_i]\}+var\{E[\bar{Y}_i|\mu_i,\sigma_i^2,n_i]\}$. From our assumptions on Y_{ij}, this gives $var\{\bar{Y}_i\}=E\{\frac{\sigma_i^2}{n_i}\}+Var\{\mu_i\}$. We are interested in finding an unbiased estimator for $var\{\mu_i\}=\sigma_{*\mu}^2$.

We observe that an unbiased estimator for $E\{\frac{\sigma_i^2}{n_i}\}$ is the sample mean of the sample variances of the means from each study, $\frac{1}{N}\sum_{i=1}^{N}\frac{s_i^2}{n_i}$, as

$$\begin{aligned} E\{\frac{1}{N}\sum_{i=1}^{N}\frac{s_i^2}{n_i}\} &= \frac{1}{N}\sum_{i=1}^{N}E\{\frac{s_i^2}{n_i}\} \\ &= \frac{1}{N}\sum_{i=1}^{N}E\{E[\frac{s_i^2}{n_i}|\mu_i,\sigma_i^2,n_i]\} \\ &= \frac{1}{N}\sum_{i=1}^{N}E\{\frac{\sigma_i^2}{n_i}\} = E\{\frac{\sigma_i^2}{n_i}\} \end{aligned}$$

An unbiased estimator for $var\{\bar{Y}_i\}$ is given by the sample variance of the means, $\frac{1}{N-1}\sum_{i=1}^{N}(\bar{Y}_i-\bar{Y}_+)^2$, where $\bar{Y}_+ = \frac{1}{N}\sum_{i=1}^{N}\bar{Y}_i$.

Therefore, an unbiased estimator for $\sigma^2_{*\mu}$ is given by $\boxed{\frac{1}{N-1}\sum_{i=1}^{N}(\bar{Y}_i-\bar{Y}_+)^2-\frac{1}{N}\left(\sum_{i=1}^{N}\frac{s_i^2}{n_i}\right)}$, where $\bar{Y}_+ = \frac{1}{N}\sum_{i=1}^{N}\bar{Y}_i$.

(b) The data provided gives an estimate of $\sigma^2_{*\mu}$ equal to 98.62, which means that between-group variance, $var\{\mu_i\}$, has a larger contribution than within-group variance, $E\{\frac{\sigma_i^2}{n_i}\}$, to $var(\bar{Y}_i)$, and it is unlikely that $var\{\mu_i\}=0$, so there is a real study-to-study difference.

We can formally test this statement $H_0 : var\{\mu_i\}=0$ by a parametric bootstrap. The p-value is small, so we reject the null hypothesis.

This question and material will not be graded and will not be covered in the exam. The purpose of this question is to expand your knowledge of the hierarchical model.

R code for calculating the estimate and conducting the hypothesis test is given below:

```
rm(list=ls())
set.seed(1)
ybar<-c(59,23,41,28,40,31,34,20,30,25)
sigsq<-c(2475,1971,2009,1771,2484,2139,2275,1411,2379,1622)
n<-c(52,48,47,163,135,150,37,111,62,100)
N<-length(n)
sigsqmu<-var(ybar)-mean(sigsq/n)
sigsqmu
## parametric bootstrap
## under H0: sigsqmu==0, mu1=mu2=...=mu
## step 1: yij~norm(mu,sigsqi)
B<-50000
ybar.boot<-rep(NA,N)
sigsq.boot<-rep(NA,N)
sigsqmu.boot<-rep(NA,B)
muhat<-mean(ybar)

for(b in 1:B){
  for(jj in 1:N){
    yij<-rnorm(n[jj],muhat,sqrt(sigsq[jj]))
```

```
    ybar.boot[jj]<-mean(yij)
    sigsq.boot[jj]<-var(yij)
  }
  sigsqmu.boot[b]<-var(ybar.boot)-mean(sigsq.boot/n)
}
hist(sigsqmu.boot)
pval=mean( abs(sigsqmu.boot)>abs(sigsqmu))
pval
#pval is 0.00018 ;therefore, we reject H0 and conclude that there is real
#study to study variability. Intuitively, this is because var(ybar) has two
#contributions sigsqmu and mean(sigsq/n).
#sigsqmu is much larger than mean(sigsq/n), and therefore,
#it is unlikely that sigsqmu is zero.

sigsqmu
var(ybar)
mean(sigsq/n)
```

Problem 4

(a) Let σ^2 be the variance in an individual's response to treatment, and let N_A and N_B denote the number of patients allocated to treatments A and B, respectively. We are trying to minimize $\sigma^2(\frac{1}{N_A} + \frac{1}{N_B})$ subject to the constraint $150N_A + 100N_B \leq 30000$.

Using the Lagrange multiplier, we want to minimize $F(N_A, N_B, \lambda) = \sigma^2(\frac{1}{N_A} + \frac{1}{N_B}) - \lambda(30000 - 150N_A - 100N_B)$.

Taking partial derivatives with respect to all parameters and setting them equal to 0 gives

$$\begin{aligned} -\frac{\sigma^2}{N_A^2} + 150\lambda &= 0 \\ -\frac{\sigma^2}{N_B^2} + 100\lambda &= 0 \\ 30000 - 150N_A - 100N_B &= 0 \end{aligned}$$

We can solve for N_B in terms of N_A in the last equation to obtain $N_B = 300 - \frac{3}{2}N_A$. We solve for λ in the first equation, plug it into the second, and substitute $300 - \frac{3}{2}N_A$ for N_B to obtain

$$\frac{3}{2}\frac{\sigma^2}{(300 - \frac{3}{2}N_A)^2} - \frac{\sigma^2}{N_A^2} = 0$$

We multiply both sides by $N_A^2(300 - \frac{3}{2}N_A)^2$ to obtain $\sigma^2(\frac{3}{2}N_A^2 - (300 - \frac{3}{2}N_A)^2) = 0$. We assume $\sigma^2 > 0$, so this only occurs if

$$\frac{3}{2}N_A^2 - (300 - \frac{3}{2}N_A)^2 = 0$$

We multiply by -4/3 and simplify to get $N_A^2 - 1200N_A - 120000 = 0$. The quadratic formula gives $N_A = 600 \pm \sqrt{600^2 - 120000} = 600 \pm 200\sqrt{6} = 110.10, 1089.90$ as our possible solutions. Because only the first one gives $N_B > 0$, we take it as our solution.

To check that this is a minimum, we check the second derivative: $\frac{\partial^2}{\partial N_A^2} F(N_A, 300 - \frac{3}{2}N_A, \lambda) = \frac{2(3/2)^2}{(300-\frac{3}{2}N_A)^3} + \frac{2}{N_A^3}$ which is positive for our value of N_A. This means that our value, about 110.10, indeed gives the minimum variance.

We cannot choose a fractional number of patients, so we test the nearest solutions $N_A = 110$ and $N_A = 111$. These give $N_B = 135$ and $N_B = 133.5$, which we round down to $N_A = 133$. These, in turn, give variances of $\sigma^2(0.01650)$ and $\sigma^2(0.01653)$.

Therefore, $\boxed{\text{the optimal allocation is to treat 110 patients with regimen A and 135 patients with regimen B}}$.

(b) The optimal treatment had a variance of $\sigma^2(0.01650)$. For an allocation with n patients on each regimen, the variance would be $\sigma^2(\frac{2}{n})$. Then we want $n \geq 2/0.01650 = 121.22$, or $n = 122$ at the minimum. This means we need to spend $122(150 + 100) - 30000 = \boxed{500}$ more dollars to get the same degree of precision as the allocation in part (a).

It is okay if set n=121. In this case, the variance would be \sigma^2(0.01652), which is slightly larger than \sigma^2(0.01650), but this difference is really inconsequential.