Sample Project – Hospital Readmissions Project Report Template

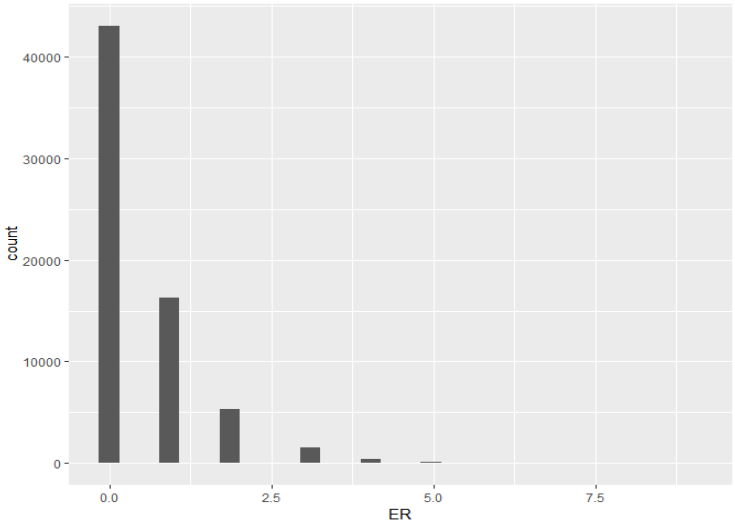
Instructions to Candidates: Please remember to avoid using your own name within this document or when naming your file. There is no limit on page count.

As indicated in the instructions, work on each task should be presented in the designated section for that task.

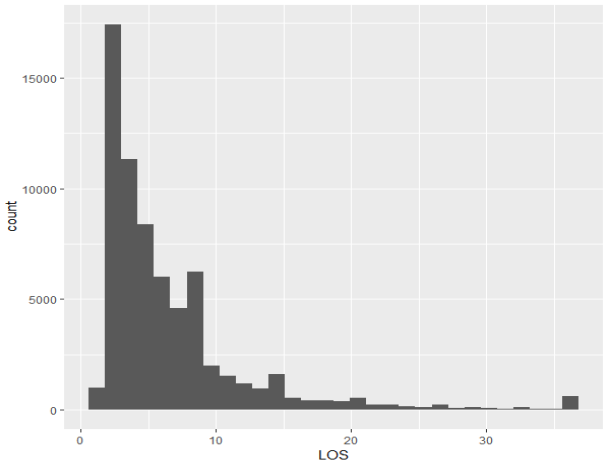
Task 1 – Perform univariate exploration of the four non-factor variables (6 points)

The four continuous variables are ER, which is the number of ER visits that the patient has had, LOS, the Length of Stay at the hospital, Age, and HCC.Riskscore, a measure of the patient's relative healthcare costs.

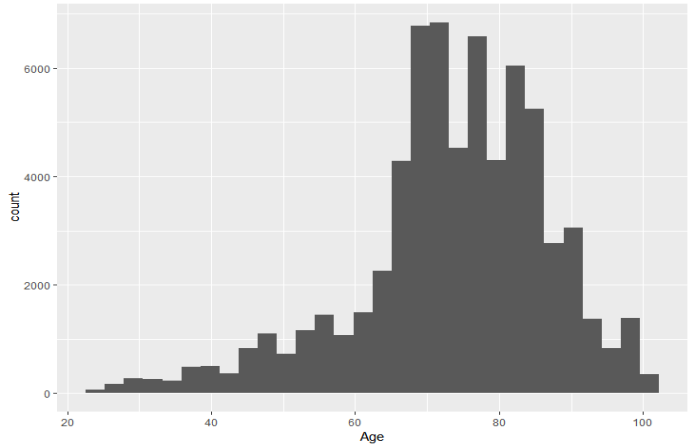
The number of ER visits ranges from 0 to 9. Most patients have not had an ER visit. As the graph below shows, this is right-skewed and so a log transform will be applied.



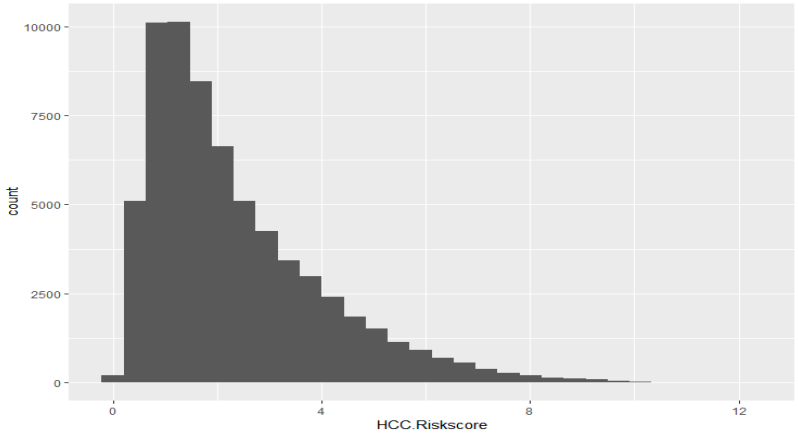
The length of stay (LOS) ranges from 1 to 36. Notice that there are many patients who have a LOS value of 36 which indicates that their stay was *at least* 36 days. A new indicator variable will be created where LOS is 36 in order to capture this. Because this is right-skewed, a log transform was applied.



The patients observed are generally over the age of 65, as this is when a person becomes eligible for Medicare. The youngest patient is 24 and the oldest is 101. The results of this study will not apply to people under the age of 24 as this is outside of the range of the data.



The field HCC.Riskscore measures the patient's healthcare costs. Patients with a higher score tend to incur greater costs. This ranges from 0.079 to 12.3 and has a median of 1.865. Because this is right-skewed, a log transform was applied.



Task 2 – Examine relationships between DRG.Class and DRG.Complication (5 points)

The DRG.Class and DRG.Complication factors are closely related. Patients that have a Class of Medical can only have medical complications and patients with a Surgical Class can only have a surgical complication. There are 6 patients which have a surgical complication but are in a medical class. These are taken to be in error and were removed.

DRG.Complication	MED	SURG	UNGROUP
MedicalMCC.CC	18104	6	NA
MedicalNoC	12310	NA	NA
Other	5357	3424	564
SurgMCC.CC	NA	15468	NA
SurgNoC	NA	11549	NA

The levels of DRG.Class and DRG.Complication were combined into a new column named DRG. The counts of each level are below. Note the high number of cases in the "other" category which are comprised of patients with a DRG Complication of Other or a DRG class of UNGROUP. These patients could be looked into in greater detail to see if they should be treated differently.

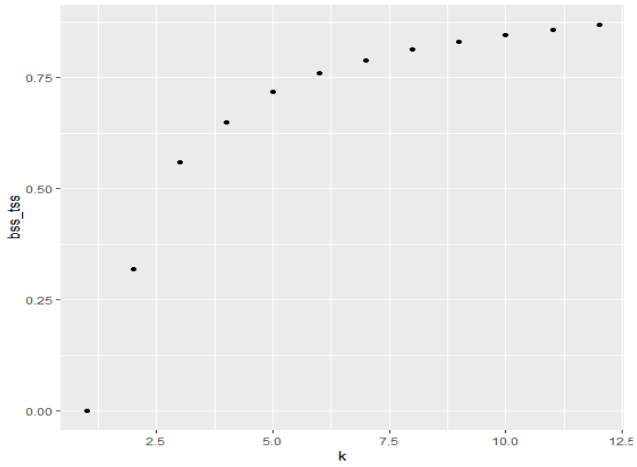
DRG	n
MedicalMCC	18104
MedicalNoC	12310
Other	9345
SurgMCC	15468
SurgNoC	11549

Task 3 – Use observations from cluster analysis to consider a new feature (9 points)

Kmeans is an unsupervised learning technique which tries to create homogenous groups of observations called clusters. It starts by first selecting random centers, which is based on a random seed value. Then it iteratively moves the cluster centers until the within-cluster variance is minimized. This is a computationally difficult task and so the algorithm runs in a greedy-forward-stage wise manner. Often times this works, but it is possible to get "stuck" at a local minimum instead of a global minimum of total within-cluster variance.

Running the algorithm multiple times with different starting positions and then taking the average is one way to reduce the risk of getting stuck like this. The nstart parameter controls the number of times that this is run. Higher values are generally better but due to computer resources anywhere from 10-100 is acceptable. I set this to 30.

The number of clusters in kmeans needs to be set manually. The way that this is generally chosen is by looking at a plot of the within-cluster deviance against the number of clusters and choosing the value of k which follows on the "elbow". The below graph shows the cluster deviance against values of k from 1 to 12. There is not a clear elbow in this graph. A value of k = 4 was selected.



Task 4 –Select an interaction (5 points)

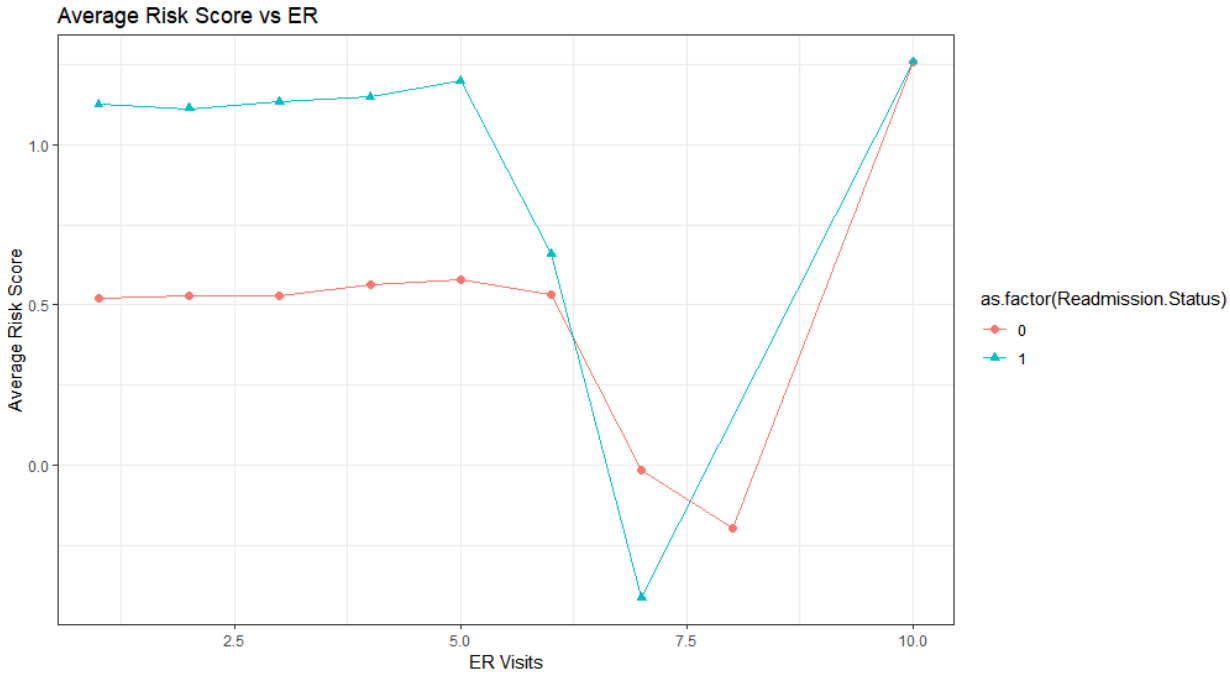
An interaction effect is where the effect of a variable on the response varies depending on the value of a different variable. Two interaction terms were considered based on the data 1) Race with Gender and 2) Riskscore with ER.

The table below shows the readmission rates by race and gender status. This shows that the percentage of readmitted patients by gender and race does appear to have an interaction. Black males have a rate of 12.9% while black females have a rate of 14.2%, whereas for white males and white females there is only a difference of 0.001.

Gender	White	Black	Hispanic	Others
1 F	0.125	0.129	0.148	0.109
2 M	0.126	0.142	0.120	0.124

While this does appear to be statistically significant, there are ethical and legal considerations to be taken into consideration when using Race in a model and so this will not be considered further.

The graph below shows the average log of the risk score against the number of ER visits for patients where readmitted (triangle) and those where were not (circle). The interaction occurs when the ER visits are between 6 and 9 where the two lines cross. This suggests that risk score has a different effect on the readmission rate depending on the number of ER visits.



Task 5 – Select a link function (8 points)

The data was split into 75% train and 25% test sets. A random seed was set to insure reproducibility. The readmission rates were approximately the same between the two data sets. There were 12.5% and 12.8% for train and test respectively.

The readmission status is either 0 or 1 which means that the binomial distribution is the appropriate choice. There are five possible link functions to use. Each will be tested in the model and compared based on AUC. For each model at this step, all variables will be included.

Logit

The logit is the default link function and also the canonical link for the binomial distribution. It maps the linear predictor to the probability space of [0,1] via the log of the odds, where the odds are the probability of readmission divided by the complement of this probability. This makes interpretation difficult. Because the objective here is to predict the most readmitted patients, this factor is not an issue.

The AUC and AIC on the test sets are below.

AUC: 0.7447
AIC: 33857

Probit

The probit is a name for the inverse cumulative normal CDF and makes interpreting the model easier by providing interpretation of the predictions as a normal probability. The AUC was slightly worse than the logit. The AIC is slightly better. The probit response was ultimately chosen although the logit would likely also be a viable option in this case.

AUC: 0.7446
AIC: 33841

Cauchit

This link function is not as popular as the probit or logit because it does not work on as broad of a class o f problems. The link function is complex, which makes interpretation more difficult. The AUC here is wo rse than the logit and probit and so it will not be considered.

AUC: 0.7443
AIC: 34381

Log

The log link does not map to [0,1] which means that the predicted probabilities would not make sense in this context. This was not tested for this reason.

Complementary Log Log

The AIC is lower than the other link functions and so this will not be used.

AUC: 0.7443
AIC: 34381

The final choice of link was the probit because it had the lowest AIC and highest AUC.

Task 6 – Decide on the factor variable from Task 3 (5 points)

In task 3, a new feature was created based on clustering the log of LOS and Age. The question here is whether to use this feature as a replacement of the original variables. Two models were fit. The first included the original variables and the second used the cluster feature instead.

With the cluster feature the AIC was 33,853 and the AUC was 0.7444. With the original variables this was 33,836 and 0.7446, which is better. The original variables were used instead of the cluster feature.

Task 7 – Select features (15 points)

To further improve the model we can remove variables which do not add statistical value. The factors DRG and Race have many levels, some of which add value and some of which do not. If StepAIC was run over these variables as they are then they would either keep all factor levels or remove the variable entirely.

Step AIC is based on the penalized log-likelihood of the response given the data. AIC adjusts the log of the likelihood by imposing the number of parameters as a penalty. When used in the backwards direction, the stepwise algorithm first begins with all variables included and then interactively removes variables which do not improve the model. When using the backwards direction, it starts with no variables and interactively adds variables which improve the model. The default in R is to run in both directions and then only keep the best variables, which is what was done here.

The solution to this is to first create binary (dummy) variables for these factor levels, and then run step AIC. Using step AIC with a backwards direction will start with all variables and factor levels as they are and then only remove the variables or factor levels, one at a time, if they do not add value to the model.

Running the step AIC in this manner resulted in a simpler model. Only DRGMedicalNoc, DRGOther, log_LOS, Age, and the log of risk score were kept. The final AIC was 33,823 and AUC of 0.7446 on the test set. Both of these metrics are better than the previous models. Based on the test set performance, the p-values are all less than 0.05, with the exception of DRGMedicalNoC which is just slightly higher at 0.068.

The AUC of 0.7446 is an improvement over the LACE index of 0.70.

Task 8 – Interpret the model (6 points)

The results from retraining the model over the entire data set are below. Note that the AIC is higher in this instance because the model was evaluated on the same data which it was trained. The advantage to retraining over the entire data set is that the coefficients are more stable (lower variance) due to a larger sample size.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4995836  0.0408970 -36.667  < 2e-16 ***
```

```
DRGMedicalNoC -0.0166168  0.0173756  -0.956 0.338905
DRGOther       0.0558087  0.0190201   2.934 0.003344 **
log_LOS        0.0333907  0.0097296   3.432 0.000599 ***
Age           -0.0036983  0.0004917  -7.522 5.39e-14 ***
log_riskscore  0.7006524  0.0107255  65.326  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 50559  on 66775  degrees of freedom
Residual deviance: 45116  on 66770  degrees of freedom
AIC: 45128

Number of Fisher Scoring iterations: 5
```

The most importance factors to whether or not a patient is readmitted to the hospital are

- Whether or not they are in a DRG medical class or have a medical complication
- Whether or not they in the “other” bucket for DRG class and complication. Refer to Task 2 where this was examined in detail.
- The length of the inpatient stay
- Their age
- Their risk score

Interpreting a logistic regression model with a probit link is not as easy as reading off the coefficients as in a regression glm. The fact that these variables have stayed in the model tells us that they are predictive, but it is not clear if they increase or decrease the readmission rate.

One way of measuring this is to consider the average patient. I first found the average patient, and then calculated the predicted readmission rate, y, based on changing the member’s characteristics.

Patients likely to be readmitted tend to

- **Have a higher risk score.** See the rate of 37.4% for the patient with the highest risk score
- **Be younger.** See rate of 13.1% vs. the first row, which is the average patient of 9%
- **Have more ER visits.** See row 4 with rate of 12.8% against average patient of 9.7%.

Gender	Age	log_LOS	log_ER	log_riskscore	RaceHispanic	RaceOthers	y
M	75	1.653	0.302	0.6	0	1	0.097
M	75	1.653	0.302	2.0	0	1	0.374
M	30	2.000	0.302	0.6	0	1	0.131
M	30	1.653	3.000	0.6	0	1	0.128

Task 9 – Set the cutoff (9 points)

The output of the model is a probability of readmission. The goal is to predict a binary outcome, whether or not the patient will need to be readmitted. A cutoff value is the threshold from 0-1 that determines if a patient is predicted to be a readmission.

The ideal cutoff value is such the patients above are segmented so that the total costs to the hospital are minimized. There are three basic scenarios

1) Do not employ intervention at all and instead pay the fee of $25 for each readmission. The cost of this would be $210,225.
2) Employ intervention to every discharges patient at a cost of $2 /patient. The total cost would be 66,776 patients * $2/patient = 133,552
3) Only intervene for the patients that the model predicts to be readmissions. I optimized this over several cutoff values from 0.01 to 0.5. For each cutoff value, the total cost was calculated as $2 for each intervention + $25 for those patients which the model missed who were actually readmitted. As you can see, this means that costs can be reduced to $105,844 which results in $27,708 in savings (21%) as compared to having an intervention for every patient.

Cutoff	Total Readmission Cost
0.01	$ 129,874
0.075	$ 106,258
0.08	$ 105,844
0.1	$ 107,824
0.2	$ 141,920
0.3	$ 180,562
0.5	$ 210,017

Setting this cutoff to 0.08 results in the following confusion matrix. About 26,000 patients, (40%) are predicted to not have a readmission and correctly do not. 1,090 (2%) patients are not predicted to but are in fact readmitted.

Prediction	Reference 0	Reference 1
0	26389	1090
1	31978	7319

Prediction	Reference 0	Reference 1
0	40%	2%
1	48%	11%

Task 10 – Consider alternative models and model construction techniques (12 points)

The objective is create a model which predicts patient readmissions with high accuracy. Interpretation is less important than performance and ease-of-implementation at the hospital. Several other modeling methods which could lead to more cost savings are

LASSO Regularized Regression

Advantages:

- Creates a simple-to-deploy linear model ;
- Performs variables selection;
- Provides consistent predictions which do not change with the underlying data as much as other methods can;
- Easier to interpret than other non-linear models;
- Increases predictive power due to L1 or L2 (in elastic Net) penalized coefficients;
- Performs well on small data sets. Not all hospitals may have 70,000 patients with which to train a model.

Disadvantages:

- Cannot handle as many link functions in R. If using Python this is not an issue;
- Sensitive to outliers (same as current model);
- Does not detect nonlinearities;
- Does not handle missing values;
- Interaction effects must be hand-coded which is time consuming and less powerful than automated methods;

Classification tree using cost-complexity pruning

Advantages:

- Easy to interpret which allows for hospitals to understand *why* certain patients get readmitted. This allows for costs to be decreased by decreasing the overall number of readmissions. Patients have are healthier as well;
- Detects interaction effects;
- Handles missing values;
- Not sensitive to outliers;
- Handles categorical and continuous data.

Disadvantages:

- High variance which results in model changing form upon being retrained. This is counter-intuitive to those relying on simple tree interpretation;
- Low predictive performance due to each prediction being the average of training data cases; each prediction is an overage of the observation is an average over the terminal node. This is a simplification of the underlying process;
- Has a bias towards selecting high-cardinality and continuous features due to variables at each split being determined by information gain (or log loss, entropy, etc) at all possible split points

Random Forest

Advantages:

- Bagging process improves predictive power;
- Results are more stable upon retraining due to lower variance ;

- Can use more features than a single tree without overfitting due to bagging of multiple trees;
- Detects interaction effects;
- Handles missing values;
- Not sensitive to outliers;
- Handles categorical and continuous data;
- Does not require log transforms due to split algorithm being based on rank order instead of variable value.

Disadvantages:

- Difficult to interpret;
- Computationally expensive to implement as compared to a decision tree which is just a set of if/else statements or a linear model which is a product of relativities. It is not possible to program a random forest inside a SQL database;
- Requires a larger training set. Not all hospitals are the same size and so the sample size needs to be taken into consideration.

Task 11 – Executive summary (20 points)

A reduction in hospital readmissions leads to 1) improved patient health outcomes and 2) cost savings from CMS readmission fees. Based on data from 66,000 patients, a predictive model was constructed which can triage high-risk patients who are likely to be readmitted. When comparing the total hospital cost to a standard intervention program that would be $2 for each patient, the cost savings is $27,708 annually.

By using predictive analytics, multiple factors are taken into consideration for each patient such as their risk score, age, gender, race, and medical profile. This information is then used to detect "red flags" among patients who are likely to need to be readmitted after leaving the hospital. This model is statistically better than the LACE index based on the AUC score (0.7442 compared with LACE of 0.70), which means that it will save the hospital more money by preventing more readmissions.

Patients who are likely to be readmitted tend to

- Have a higher risk score;
- Be younger;
- Have more ER visits.

The most important factors to consider are

- Whether or not they are in a DRG medical class or have a medical complication
- Whether or not they are in an unusual surgical or medical DRG complication and medical status (see the "Other" category in Task 3).
- The length of the inpatient stay
- Their age
- Their risk score

This analysis is based on 66,000 hospital patients and data on whether or not they have been readmitted, the risk score, age, gender, race, number of previous ER visits, DRG class, and Length of Stay (LOS). Minor adjustments were made to the data such as taking the logarithms of ER visits, LOS, and risk score. Six patients with incorrect DRG records were removed. Additional factors were taken into consideration such as if their LOS was greater than 36 or if they had an interaction effect between risk score and ER visits, and different groupings of LOS and age (clustering).

There was overlapping information between two of the variables, DRG.Complication, and DRG.Class. To simplify, these were combined into a single variable. Factors which were not seen to be statistically significant were the patient's race, certain DRG classes which were related to medical or surgical classes, and gender. The Race variable was remove due to potential ethical and legal complications.

To ensure that the model will work as well in real life as it has in this analysis, a holdout set of 25% of the patients was created. The model did not "see" these patients while it was learning from the data. Several models were tested using different statistical methods and combinations of inputs. Each of these were evaluated based on how well they predicted readmissions in the "unseen" data and then the best model was chosen from these.

The model was fine-tuned to minimize the readmission cost using scenario testing. Three scenarios were considered

1) Do not employ intervention at all and instead pay the fee of $25 for each readmission. The cost of this would be $210,225.
2) Employ intervention to every discharged patient at a cost of $2 /patient. The total cost would be 66,776 patients * $2/patient = $133,552
3) Only intervene for the patients that the model predicts to be readmissions. This results in the lowest cost of **$**105,844, or in $27,708 in savings (21%) as compared to having an intervention for every patient.

This final model is based on the patient's risk score, which I understand to be an estimated quantity of future medical costs. If this risk score were to change then the model would need to be updated.

The patients in this study were older, comprised mostly of Medicare patients over 65. The youngest patient was 24. This means that the model can only detect readmissions for this age group of patients. If the results need to apply to a younger population, then new data should be supplied with these young patients.

The comparison with the LACE index is based on the AUC (Area Under the Curve) which is a common statistical metric. In order to get a more precise measure of the improvement of the model as compared to LACE, both models would need to be tested on similar populations and sample sizes.