Notes on combining p-values of GWAS studies from different populations

Ziang Zhang

1 Introduction:

For most human traits, genetic effects from specific SNP only have small effect sizes (Evangelou and Ioannidis 2013). Therefore, practitioners often aggregate the GWAS results from different populations through meta-analysis methods in order to achieve higher power. Aggregation through p-values is a commonly used example of meta-analysis method. For example when there are several different datasets, practitioners sometimes will first carry out a comprehensive study using one dataset, and only follow up with the SNPs that have the highest significance levels (Begum et al. 2012). In this note, we would like to consider whether such procedures will be appropriate for the analysis of binary traits.

With examples of Wald test, we presented the phenomenon that the same hypothesis test on different populations can have very different p-values under the alternative hypothesis, hence very different powers, even if there is no heterogeneity on the true SNP effect and the covariates among the two datasets are identical.

In next section, we will give a brief overview of Wald tests for Generalized Linear Models, with a short explanation of the rationale behind this phenomenon. Then, we will follow with two simulation studies to illustrate how this phenomenon can occur for binary trait.

2 Wald test for Generalized Linear Models:

Assume the generalized linear model (glm) has the following form:

$$\mathbb{E}(Y|X) = \mu = g^{-1}(\beta_0 + \beta_1 X_1 + \beta_2 X_2) = g^{-1}(\eta),$$

where $g(.)$ is a specific link function connecting the linear predictor η with the mean function of Y. In this case, the fisher information matrix at β can be written as

$$I_n(\beta) = XW(\beta)X^T,$$

where X denotes the design matrix and $W(\beta)$ is a diagonal matrix with each diagonal term depending on the value of β unless g is identity function. Specifically, the i^{th} diagonal term of W can be computed as

$$w_i = (\frac{\partial u_i}{\partial \eta_i})^2 / \mathrm{Var}(Y_i|X).$$

If the question of interest is to test the hypothesis $H_0 : \beta_2 = 0$ using Wald test, the test statistic can be written as

$$T = I_n^{-1}(\hat{\beta})_{[3,3]}(\hat{\beta}_2)^2,$$

where $I_n^{-1}(\hat{\beta})_{[3,3]}$ denotes the third diagonal term of the matrix $I_n^{-1}(\hat{\beta})$ and $\hat{\beta}$ is the MLE estimator. Under the null hypothesis, T asymptotically follows a Chi-Square distribution with 1 degree of freedom.

Under the alternative hypothesis that $\beta_2 = \tilde{\beta}_2 \neq 0$, the non-centrality parameter of Wald test above can be computed as

$$I_n^{-1}(\tilde{\beta})_{[3,3]}\tilde{\beta}_2^2,$$

where $\tilde{\beta}$ is the vector of true values for the regression parameters β. Since $I_n^{-1}(\tilde{\beta})_{[3,3]}$ will not solely depend on $\tilde{\beta}_2$ unless g is identity. Therefore the power function of this Wald test will not only depend on $\tilde{\beta}_2$, but the whole vector $\tilde{\beta}$.

Define $d = -\sqrt{I_n^{-1}(\tilde{\beta})_{[3,3]}}\tilde{\beta}_3$, the theoretical power of this Wald test at $\tilde{\beta}$ can be computed as

$$1 - \Phi(d + z_{a/2}) + \Phi(d - z_{a/2}),$$

where Φ is the CDF of standard normal and $z_{a/2}$ is the $a/2$ quantile of standard normal.

In summary, this means that if we utilize Poisson regression to analyze count traits (e.g. number of cancers) or Logistic/Probit regression to analyze binary traits (e.g. disease status), powers of Wald test from different populations can be dramatically different, even if the two populations share the same effect size (i.e. $|\tilde{\beta}_2|$) and the same set of covariates $\{X_{1i}, X_{2i}\}_{i=1}^n$. The rationale behind this is actually the classical contrast between **statistical significance** measured by p-values and **practical significance** measured by the size of the effect $|\tilde{\beta}_2|$. In Wald test, p-values are determined by both the practical significance $|\tilde{\beta}_2|$ and the standard error $\sqrt{I_n^{-1}(\tilde{\beta})_{[3,3]}}$. Since the standard errors of the MLE estimator will be different on the two populations, the conclusion drawn from statistical significance may be inconsistent with the practical significance of effects in the two populations.

3 Simulation with random samples:

Consider that two samples of size $n = 1000$ have been collected independently on two populations (European, Asian), and question of interest is to study the association between the status of a particular disease ($Y : 0/1$) and a particular SNP G with minor allele frequency (MAF) 0.3 under Hardy Weinberg Equilibrium (HWE), after controlling the effect of a covariate Z ($Z \sim N(0, \sigma = 3)$).

Assume that GWAS study has been carried out for each population to test if the SNP G is causal for Y, and the p-values are 5×10^{-4} for European population and 5×10^{-8} for Asian population, can we interpret the result as there is likely a stronger association in Asian population than European population? To answer this question, we will conduct the following simulation study.

We assume that the models that generate the observations for European population and for Asian population are the followings:

$$\textbf{Euro} : \text{logit}(\text{P}(Y = 1|G, Z)) = -0.5 + 0.8Z + 0.3G,$$

$$\textbf{Asian} : \text{logit}(\text{P}(Y = 1|G, Z)) = -0.5 + 0.1Z + 0.3G.$$

In other words, the effect of Z will be different across the two populations, but the effect of the SNP will be the same. We generated $\{G_i\}_n$ and $\{Z_i\}_n$ independently, and used the same set of covariates to generate the response variables in each population.

3.1 Summary of simulated data and P values:

In our simulation, the same sets of G and Z will be used to simulate traits in both populations. We simulated $\{G_i\}$ independently from $\{Z_i\}$, and then simulated the disease status Y using G and Z. Recall that $Z \sim N(0, 3)$ and MAF of G is 0.3. The summary of our generated data is present at below:

```r
### Simulated the common Z and G
set.seed(100,sample.kind = "Rounding")
N <- 1000
G <- sample(c(0,1,2),size = N, replace = T, prob = c(0.49,0.42,0.09))
Z <- rnorm(N,sd = 3)


### Simulate each population's disease status based on Z and G
## Eur:
beta0 <- -0.5
betaZ <- 0.8
betaG <- 0.3
ylat_Eur <- beta0 + betaG*G + betaZ*Z + rlogis(N)
y_Eur <- ifelse(ylat_Eur >=0, 1, 0)

## Asia:
beta0 <- -0.5
betaZ <- 0.1
betaG <- 0.3
ylat_As <- beta0 + betaG*G + betaZ*Z + rlogis(N)
y_As <- ifelse(ylat_As >=0, 1, 0)



### Case control counts across populations:
t <- rbind(table(y_Eur),table(y_As)) %>% as_tibble()
rownames(t) <- c("Euro","Asia")
kableExtra::kable(t, caption = "Case Control Counts across populations") %>%
  kable_styling(latex_options = "HOLD_position", font_size = 10)
```

Table 1: Case Control Counts across populations

	0	1
Euro	522	478
Asia	573	427

```r
### Case control ratio across genotypes:
t <- cbind(c(y_Eur,y_As),c(G,G)) %>% as_tibble()
colnames(t) <- c("Y","G")
t <- t %>% group_by(G) %>% summarise(ratio = sum(Y)/n())
kableExtra::kable(t, caption = "Case Control Ratio across genotypes") %>%
  kable_styling(latex_options = "HOLD_position", font_size = 10)
```

Table 2: Case Control Ratio across genotypes

G	ratio
0	0.4238901
1	0.4683841
2	0.5200000

Based on the two tables above, we can notice that the case to control ratios are similar across the two

populations. Furthermore, as we can expect, the genotypic group with more copies of the minor allele has higher case to control ratio. In our simulation, the samples are randomly collected from the two populations. This strategy may not be appropriate if the disease prevalence is very low, in which samples collected in a case-control design will be preferred. We will consider the same problem for case-control study later, and for now just focus on the case when the sample is random.

We can use Wald test to test the hypothesis $\beta_G = 0$ (i.e. G is a casual SNP) in each population:

```
## EU:
mod_Eur <- glm(y_Eur~Z + G, family = binomial(link = "logit"))
summary(mod_Eur)$coefficients[3,4]
```

```
## [1] 0.01510122
```

```
## Asian:
mod_As <- glm(y_As~ Z + G, family = binomial(link = "logit"))
summary(mod_As)$coefficients[3,4]
```

```
## [1] 0.008987236
```

Note that the p-values are 0.015 for European population, and 0.009 for Asian population. It is typically expected that for the population with smaller p-value, the magnitude of the association (i.e. $|\tilde{\beta}_G|$) should be larger. However, in this simulation example the true value of β_G is $\tilde{\beta}_G = 0.3$ for both populations, and even the covariates are exactly the same.

3.2 Theoretical Power and Empirical Power:

For now, assume that the hypothesis $\beta_G = 0$ will be tested using Wald test with $\alpha = 0.05$, then we can compute the theoretical powers of the two Wald test using the simulated data $\{G_i, Z_i\}_n$ and the true parameters vectors $\tilde{\beta}_{Euro}, \tilde{\beta}_{Asia}$:

```
## Euro:
beta0 <- -0.5
betaZ <- 0.8
betaG <- 0.3
### Theoretical Power
mod_Eur <- glm(y_Eur~Z + G, family = binomial(link = "logit"))
#### Get the design matrix:
X <- cbind(rep(1,N),mod_Eur$model[,-1])
### Compute the weight matrix W:
beta <- c(beta0,betaZ,betaG)
#beta <- as.numeric(mod_Eur$coefficients)
w <- c()
for (i in 1:N) {
  si <- as.numeric(as.numeric(X[i,]) %*% beta)
  w[i] <- (dlogis(si)^2)/(plogis(si)*(1-plogis(si)))
}
I <- as.matrix(t(X)) %*% diag(w,nrow = N,ncol = N) %*% as.matrix(X)
#### Invert to get the true covariance matrix
V <- solve(I)
### Compute the power function:
delta <- sqrt(1/V[3,3])*(0-beta[3])
```

```r
alpha <- 0.05
Power_EU <- 1- pnorm(delta - qnorm(alpha/2)) + pnorm(delta + qnorm(alpha/2))
Power_EU
```

```
## [1] 0.6193771
```

```r
## Asia:
beta0 <- -0.5
betaZ <- 0.1
betaG <- 0.3
### Theoretical Power
mod_As <- glm(y_As~ Z + G, family = binomial(link = "logit"))
#### Get the design matrix:
X <- cbind(rep(1,N),mod_As$model[,-1])
### Compute the weight matrix W:
beta <- c(beta0,betaZ,betaG)
#beta <- as.numeric(mod_As$coefficients)
w <- c()
for (i in 1:N) {
  si <- as.numeric(as.numeric(X[i,]) %*% beta)
  w[i] <- (dlogis(si)^2)/(plogis(si)*(1-plogis(si)))
}
I <- as.matrix(t(X)) %*% diag(w,nrow = N,ncol = N) %*% as.matrix(X)
#### Invert to get the true covariance matrix
V <- solve(I)
### Compute the power function:
delta <- sqrt(1/V[3,3])*(0-beta[3])
alpha <- 0.05
Power_AS <- 1- pnorm(delta - qnorm(alpha/2)) + pnorm(delta + qnorm(alpha/2))
Power_AS
```

```
## [1] 0.8618099
```

Based on the results above, we know in this simulation study, the power of Wald test will be 0.619 for the European population, and 0.861 for the Asian population. Note that Wald test on Asian population has quite larger power compared to on European population, despite the fact that the two samples are generated with same $\beta_G = 0.3$ and generated by the same set of $\{G_i, Z_i\}_n$. This suggests the p-values of Wald test may have very different distributions on the two populations. We can double check that our theoretical powers for both tests are correct using empirical powers:

To compute the empirical powers, we re-simulated the disease status in each population for $K = 800$ times, and compute the 800 p-values in each population:

```r
set.seed(100,sample.kind = "Rounding")
## Euro:
beta0 <- -0.5
betaZ <- 0.8
betaG <- 0.3
p1 <- c()
for (i in 1:800) {
  ylat_Eur_rep <- beta0 + betaG*G + betaZ*Z + rlogis(N)
  y_Eur_rep <- ifelse(ylat_Eur_rep >=0, 1, 0)
  mod <- glm(y_Eur_rep~Z+G, family = binomial(link = "logit"))
```

```
  p1[i] <- summary(mod)$coefficient[3,4]
}
emp_power <- mean(p1 <= alpha)
emp_power
```

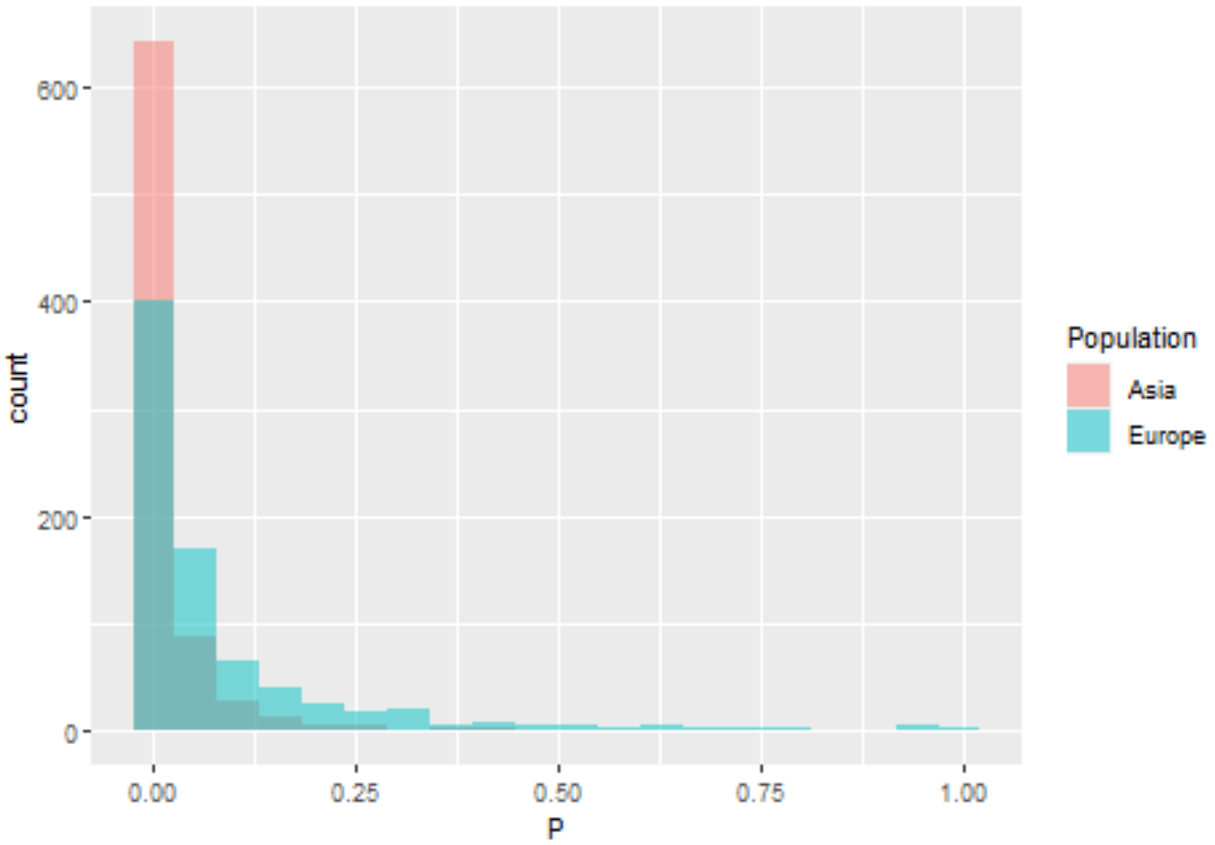
```
## [1] 0.61625
```

```
set.seed(100,sample.kind = "Rounding")
## Asia:
beta0 <- -0.5
betaZ <- 0.1
betaG <- 0.3
p2 <- c()
for (i in 1:800) {
  ylat_As_rep <- beta0 + betaG*G + betaZ*Z + rlogis(N)
  y_As_rep <- ifelse(ylat_As_rep >=0, 1, 0)
  mod <- glm(y_As_rep~Z+G, family = binomial(link = "logit"))
  p2[i] <- summary(mod)$coefficient[3,4]
}
emp_power <- mean(p2 <= alpha)
emp_power
```

```
## [1] 0.8725
```

Based on the 800 resampling results, the empirical powers are respectively 0.616 for European population and 0.873 for Asian population. These values are quite close to the theoretical values 0.619 and 0.861 we computed above. The distributions of p-values in each population can be visualized as well:

```
### Comparison:
pcomp <- tibble(P = c(p1,p2), Population = c(rep("Europe",800),rep("Asia",800)))
pcomp %>% ggplot(aes(x = P, fill = Population)) + geom_histogram(bins = 20, alpha=0.5, position="identi
```



Based on the figure above, we can conclude that the distribution of p values in Asian population is stochastically smaller than the distribution in European population, even if their underlying β_G are both 0.3. Therefore, it shows that the magnitudes of p-values of different studies are not directly comparable, unless the generalized linear regression model being used is the ordinary linear regression model with g being identity function.

4 Simulation with Case-Control design:

In this section, we will demonstrate the same problem will also occur for studies with Case-Control design, using a new simulation example.

For the new simulation, we will continue to use the two data-generating models as in section 2. However, instead of using samples randomly simulated from the population, we are now randomly sampling from cases and controls in the following way:

- First, use the two true models to generate two **populations** of size 3000 in the same way as in section 2.
- Secondly, among all the cases and controls in each population, randomly sample 500 observations from cases and 500 observations from controls.
- Finally, for each population, combine all the sampled cases and controls to have a sample with size 1000.

```
### Simulated the common Z and G
set.seed(100,sample.kind = "Rounding")
N <- 5000
G <- sample(c(0,1,2),size = N, replace = T, prob = c(0.49,0.42,0.09))
Z <- rnorm(N,sd = 3)
## Eur:
beta0 <- -0.5
betaZ <- 0.8
betaG <- 0.3
ylat_Eur <- beta0 + betaG*G + betaZ*Z + rlogis(N)
y_Eur <- ifelse(ylat_Eur >= 0, 1, 0)
Eur_controls <- tibble(Y = y_Eur, G = G, Z = Z) %>% filter(Y == 0) %>% sample_n(500)
Eur_cases <- tibble(Y = y_Eur, G = G, Z = Z) %>% filter(Y == 1) %>% sample_n(500)
EU_sample <- rbind(Eur_cases,Eur_controls)
mod_Eur <- glm(Y ~ Z + G,data = EU_sample, family = binomial(link = "logit"))
summary(mod_Eur)$coefficients[3,4]
```

```
## [1] 0.0001973326
```

```
## Asia:
set.seed(100,sample.kind = "Rounding")
beta0 <- -0.5
betaZ <- 0.1
betaG <- 0.3
ylat_As <- beta0 + betaG*G + betaZ*Z + rlogis(N)
y_As <- ifelse(ylat_As >= 0, 1, 0)
As_controls <- tibble(Y = y_As, G = G, Z = Z) %>% filter(Y == 0) %>% sample_n(500)
As_cases <- tibble(Y = y_As, G = G, Z = Z) %>% filter(Y == 1) %>% sample_n(500)
As_sample <- rbind(As_cases,As_controls)
mod_As <- glm(Y ~ Z + G,data = As_sample, family = binomial(link = "logit"))
summary(mod_As)$coefficients[3,4]
```

```
## [1] 3.561277e-33
```

Using Wald test as before, we have p-value (5.679×10^{-26}) in Asian population being much less than the p-value in European population (7.489×10^{-3}). Note that we are still assuming the same $\tilde{\beta}_G$ for each population in this example.

For case-control data, we cannot directly compute the theoretical power using the formula from section 1. When a logistic regression model is fitted for case-control data, the estimated intercept parameter $\hat{\beta}_0$ will no longer have the same interpretation as the β_0 we used in the generating model. This implies that we cannot plug in $\tilde{\beta}_0 = -0.5$ as the true value for the intercept when we compute theoretical powers. However, we can still obtain the empirical powers and distribution of p-values using the same approach as before:

```
### Simulate each population's disease status based on Z and G
## Eur:
set.seed(100,sample.kind = "Rounding")
beta0 <- -0.5
betaZ <- 0.8
betaG <- 0.3
p1 <- c()
for (i in 1:800) {
  ylat_Eur <- beta0 + betaG*G + betaZ*Z + rlogis(N)
  y_Eur <- ifelse(ylat_Eur >=0, 1, 0)
  Eur_controls <- tibble(Y = y_Eur, G = G, Z = Z) %>% filter(Y == 0) %>% sample_n(500)
  Eur_cases <- tibble(Y = y_Eur, G = G, Z = Z) %>% filter(Y == 1) %>% sample_n(500)
  EU_sample <- rbind(Eur_cases,Eur_controls)
  mod_Eur <- glm(Y~Z + G,data = EU_sample, family = binomial(link = "logit"))
  p1[i] <- summary(mod_Eur)$coefficients[3,4]
}
emp_power <- mean(p1 <= alpha)
emp_power
```

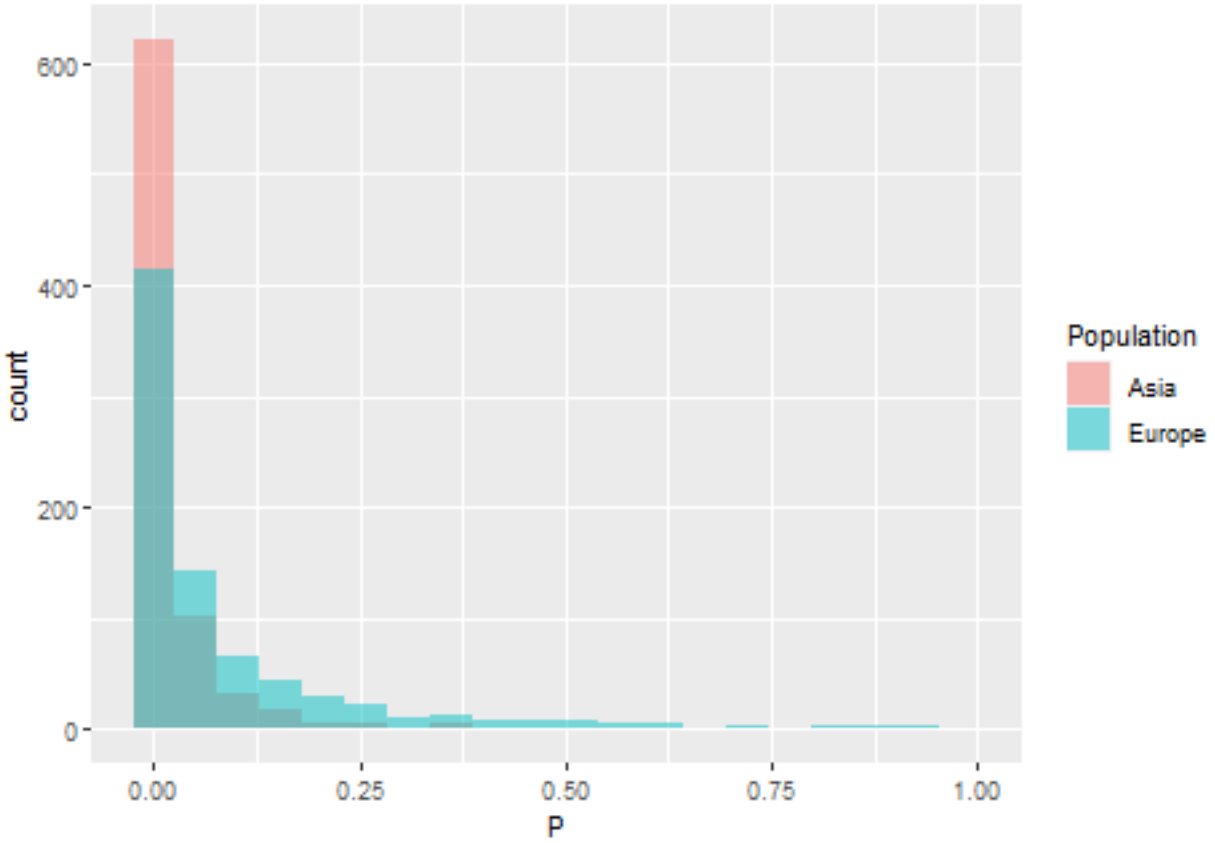
```
## [1] 0.62125
```

```
## Asia:
set.seed(100,sample.kind = "Rounding")
beta0 <- -0.5
betaZ <- 0.1
betaG <- 0.3
p2 <- c()
for (i in 1:800) {
  ylat_As <- beta0 + betaG*G + betaZ*Z + rlogis(N)
  y_As <- ifelse(ylat_As >=0, 1, 0)
  As_controls <- tibble(Y = y_As, G = G, Z = Z) %>% filter(Y == 0) %>% sample_n(500)
  As_cases <- tibble(Y = y_As, G = G, Z = Z) %>% filter(Y == 1) %>% sample_n(500)
  As_sample <- rbind(As_cases,As_controls)
  mod_As <- glm(Y~Z + G,data = As_sample, family = binomial(link = "logit"))
  p2[i] <- summary(mod_As)$coefficients[3,4]
}
emp_power <- mean(p2 <= alpha)
emp_power
```

```
## [1] 0.865
```

Again, as we have observed in section 2, the empirical powers of Wald test still differ a lot when we use case-control design (0.621 in Europe and 0.865 in Asia). We can also see the distribution of p-values for each population:

```
pcomp <- tibble(P = c(p1,p2), Population = c(rep("Europe",800),rep("Asia",800)))
pcomp %>% ggplot(aes(x = P, fill = Population)) + geom_histogram(bins = 20, alpha=0.5, position="identi
```



As we can see, the conclusion from section 2 still applies to this scenario. The rationale behind these two simulation examples is that, statistical significance (i.e. P value) does not give any information on the **practical significance** (i.e. $|\tilde{\beta}_G|$). For both populations, the practical significance of parameter β_G is the same (0.3). However, testing in the Asian population gives higher statistical significance (smaller p values), due to the larger standard error caused by the higher β_Z.

Therefore, screening out SNPs based on the significance from study in the European population might miss this casual SNP, but screening based on the Asian population likely will not. Extra caution should be given when practitioners are trying to aggregate information between two such populations.

5 Bibliography

Begum, Ferdouse, Debashis Ghosh, George C. Tseng, and Eleanor Feingold. 2012. "Comprehensive literature review and statistical considerations for GWAS meta-analysis." *Nucleic Acids Research* 40 (9): 3777–84. https://doi.org/10.1093/nar/gkr1255.

Evangelou, Evangelos, and John Ioannidis. 2013. "Evangelou E, Ioannidis Jp.meta-Analysis Methods for Genome-Wide Association Studies and Beyond. Nat Rev Genet 14:379-389." *Nature Reviews. Genetics* 14 (May). https://doi.org/10.1038/nrg3472.