Polygenic Risk Score (PRS) Introduction 101

GWAS, h^2 and prediction as the foundation

Drs. Lei Sun, Wei Deng, Yanyan Zhao

Department of Statistical Sciences, FAS

Division of Biostatistics, DLSPH

University of Toronto

10 October, 2021

At the end of this lecture, a **deeper** understanding of

- the multiple hypothesis testing issue inherent in GWAS
- the (high) variability inherent in $\hat{\beta}$, the β estimates
- heritability h^2 as a function of both β and MAF (and σ^2)
- the 'genetic effect size' of a SNP $= \beta^2 \cdot \text{MAF} \cdot (1 - \text{MAF})$
- a conceptual PRS construction based on the ground truth, PRS.oracle
- DIY ROC plotting and AUC calculation for a PRS-based prediction

GWAS is the foundation of PRS, providing J and $\hat{\beta}_j$ in $PRS_i = \sum_{j=1}^{J} \hat{\beta}_j G_{ij}$

$$Y \text{ (phenotype)} = \beta_0 + \beta_j G_j \text{ (genotype)} + \beta_E E \text{ (envir.)} + e \text{ (error)},$$
$$H_0 : \beta_j = 0,$$

where $e \sim N(0, \sigma^2)$, $j = 1 \ldots > 10^6$ **for all SNPs across the genome**.
(Could be more complex: multiple $E's$ and G's, GxE, and GxG interactions)

G_j: Genotype of a (bi-allelic, autosomal) SNP j

- coded 0, 1 and 2 for aa, Aa and AA
- a = the reference allele
- A = the alternative allele (often the minor allele with MAF of p)
- freq. of aa, Aa and AA: $(1-p)^2$, $2p(1-p)$ and p^2 under HWE

GWAS Paper '0'

WTCCC (2007). *Nature*. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.

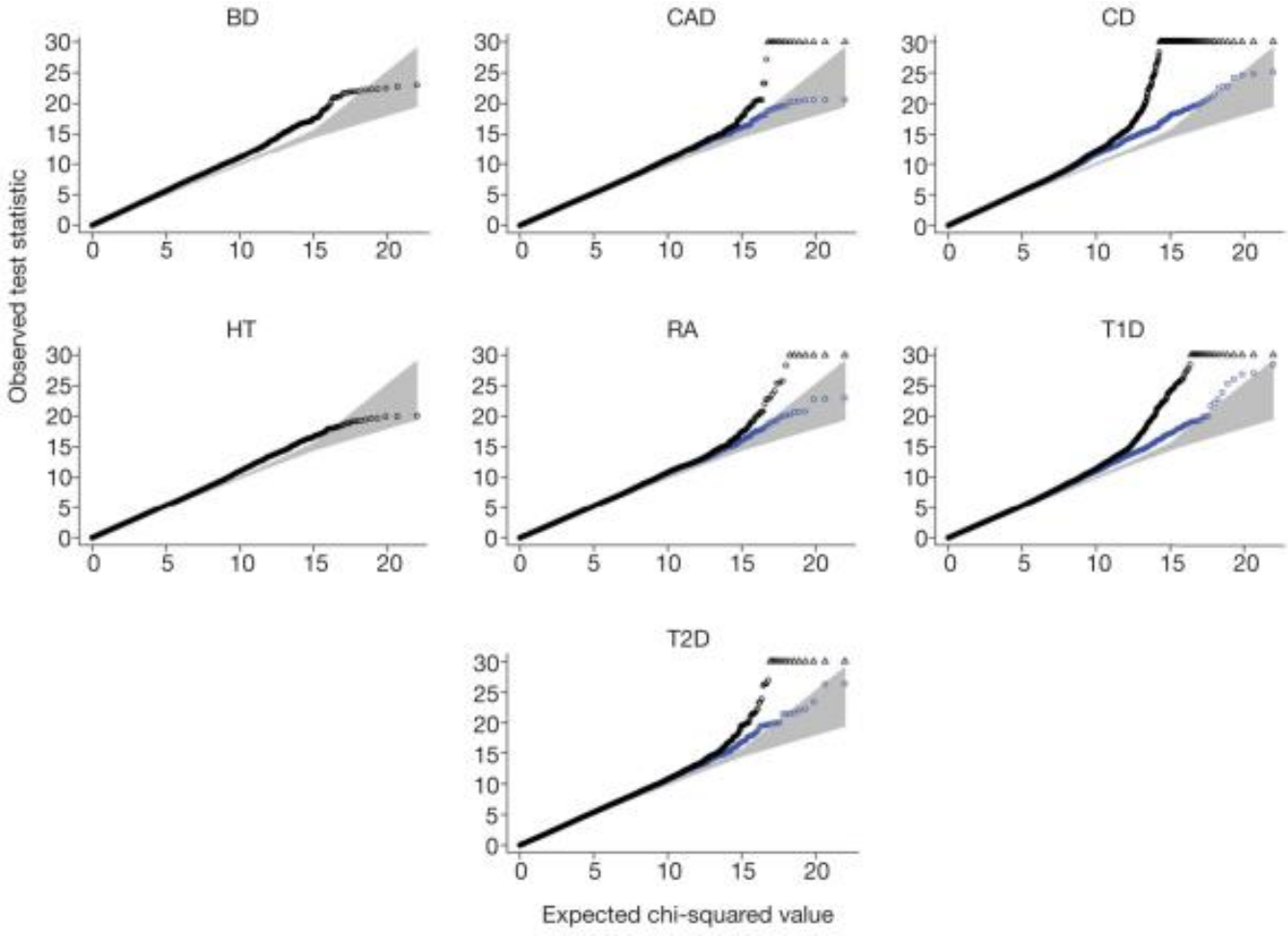
- Phenotypes: Seven major diseases, e.g. Bipolar, Hypertension
- Samples: $\approx$ 2000 cases and (shared) 3000 controls for each disease
- SNPs: Affymetrix 500K
- Analyses: **much effort on quality control (QC)**, simple association tests, novel imputation method.
- Results: 24 independent association signals at p-value $< 5 \cdot 10^{-7}$

  almost all true positives based on previous or replication studies

  Some of the loci confer risk for multiple diseases

  58 additional loci at $10^{-5} <$ p-value $< 5 \cdot 10^{-7}$

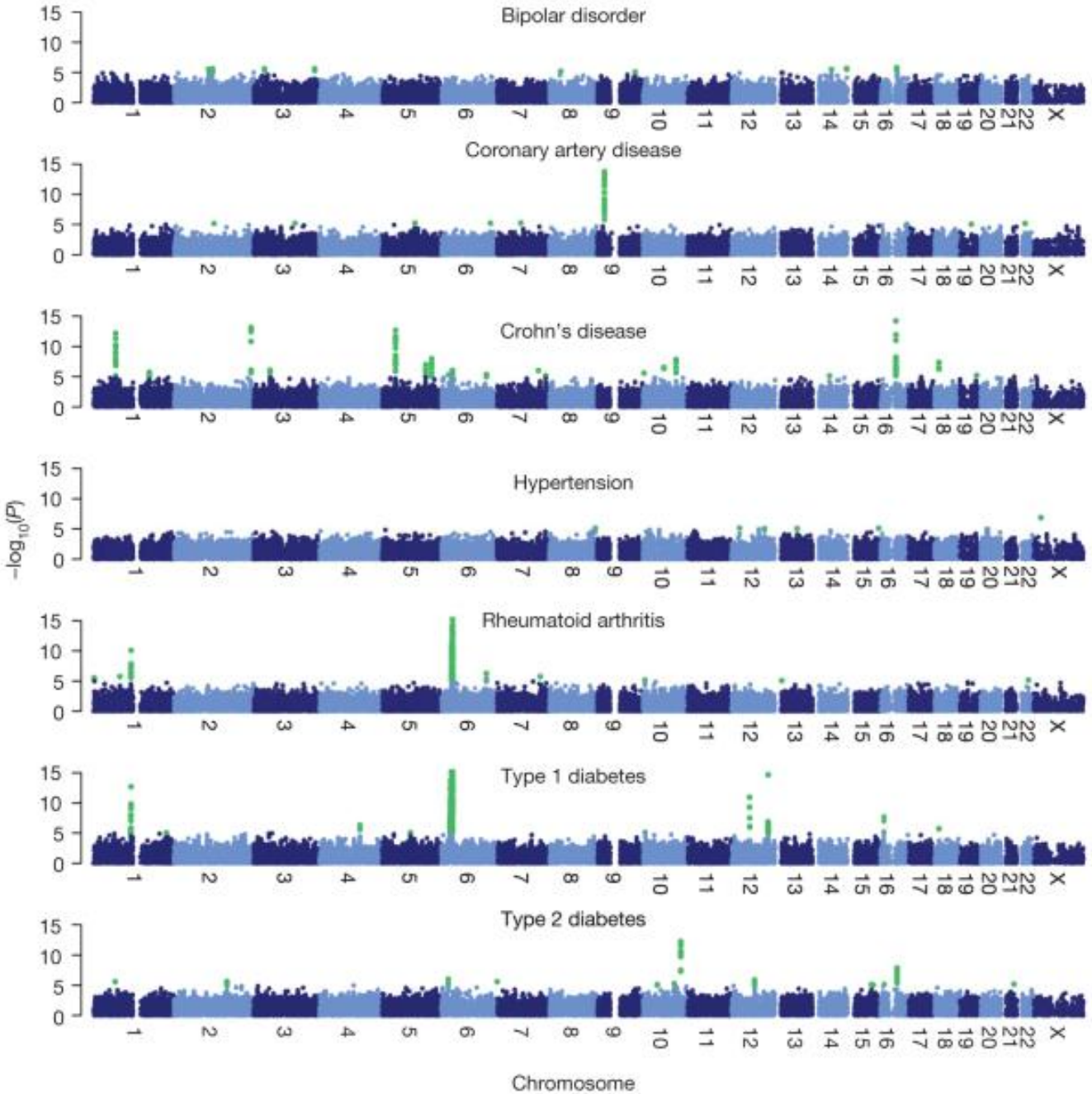
QQ-plot, Figure 3 of WTCCC 2007



Black: post-QC SNPs, MAF >1% and missing data rate <1%. SNPs at which the test statistic exceeds 30 are represented by triangles. (Most current GWASs: on the -log10(p-value) scale with no confidence band but with a main diagonal line.)

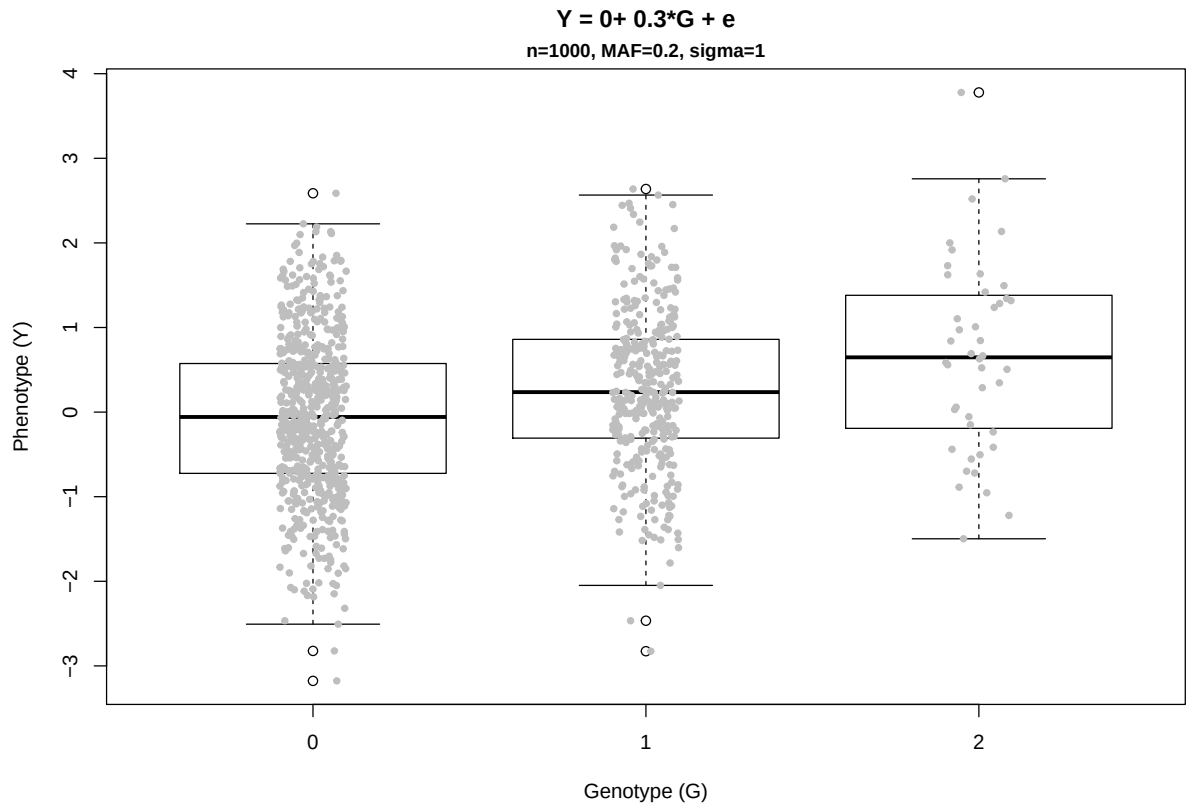
Blue: excluding SNPs located in the regions of association listed in Table 3 ($< 5 \cdot 10^{-7}$) (for BD: no visible effect on the plot, and for HT: no such SNPs).

Manhattan plot, Figure 4 of WTCCC 2007



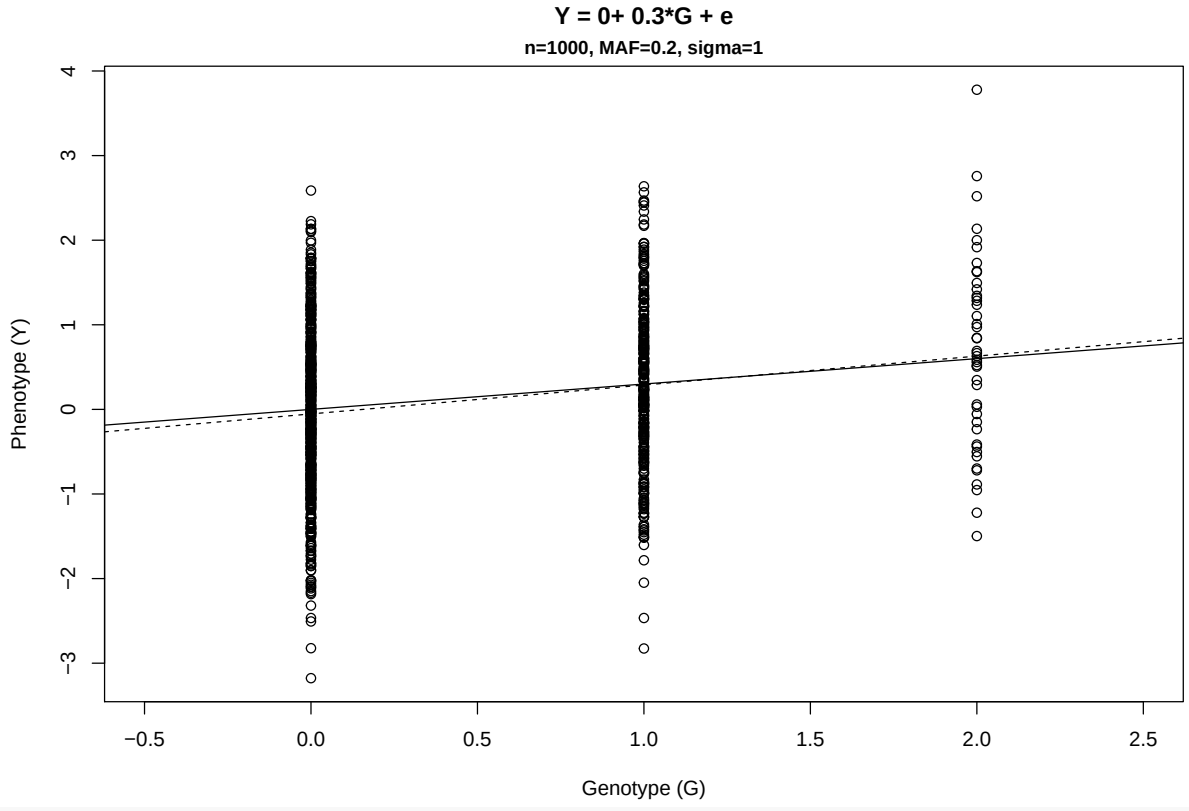
A refresher Y-on-G association test via simulation

```
set.seed(101)

nsample=1000; maf=0.2; beta=0.3; beta.0=0; sigma=1 # no E for simplicity
nG=rmultinom(1,size=nsample,prob=c((1-maf)^2,2*maf*(1-maf), maf^2)) # assume HWE
G=c(rep(0,nG[1]),rep(1,nG[2]),rep(2,nG[3]))
e=rnorm(nsample,mean=0,sd=sigma)
Y=beta.0+beta*G+e

boxplot(Y~G,main=paste("Y = ",beta.0, "+ ",beta,"*G + e",sep=""), ylab="Phenotype (Y)", xlab="Genotype (G)
stripchart(Y~G,vertical=T,method="jitter",add=T,pch=20,col="gray")
title(line=0.5,paste("n=",nsample,", MAF=",maf,", sigma=",sigma, sep=""),cex.main=0.9)
```



The true (solid) and fitted (dotted) regression lines

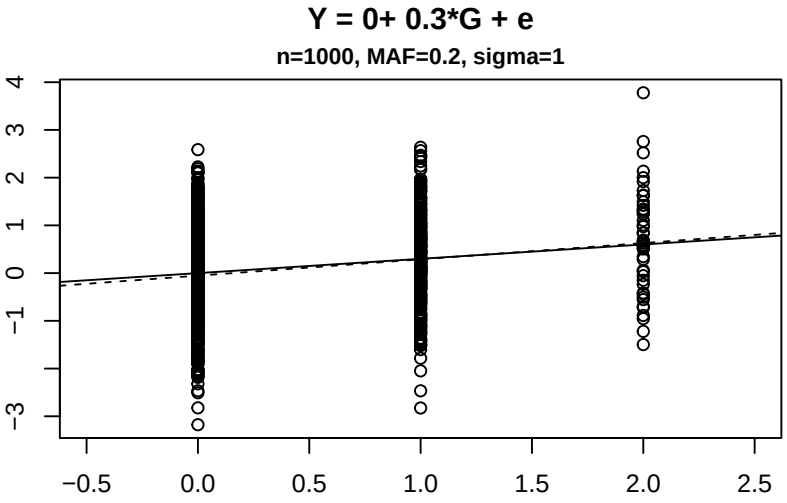
```
plot(G,Y,main=paste("Y = ",beta.0, "+ ",beta,"*G + e",sep=""), ylab="Phenotype (Y)", xlab="Genotype (G)",
title(line=0.5,paste("n=",nsample,", MAF=",maf,", sigma=",sigma, sep=""),cex.main=0.9)
fit=lm(Y~G)
abline(a=fit$coef[1],b=fit$coef[2],lty=2) # fitted regression (dotted) line
abline(a=beta.0,b=beta) # true regression (solid) line
```



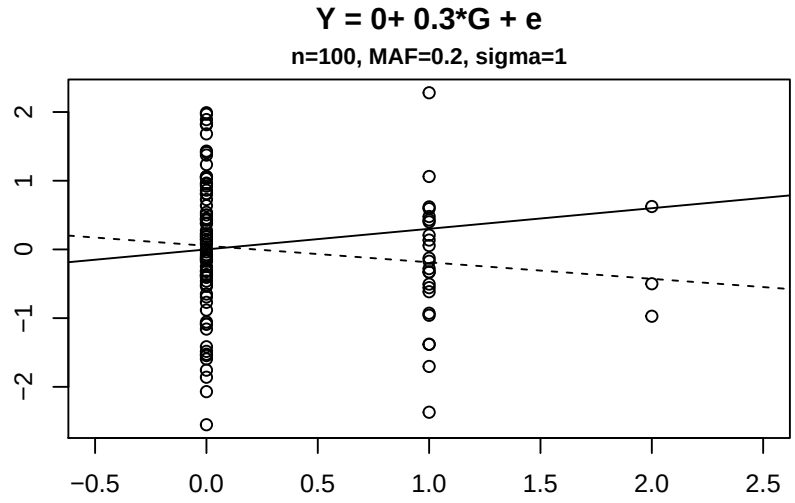

```
fit$coef
```

```
## (Intercept)           G
## -0.05350324  0.34152970
```

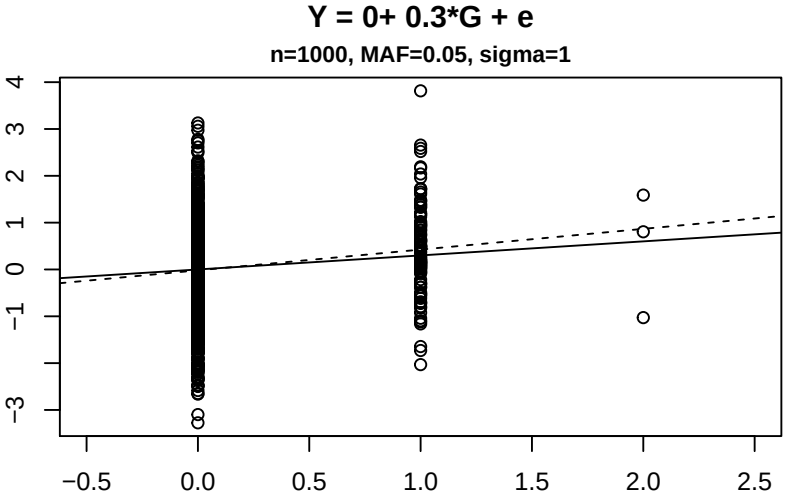
Quiz: difference between the same Y=0+0.3*G+e regression?



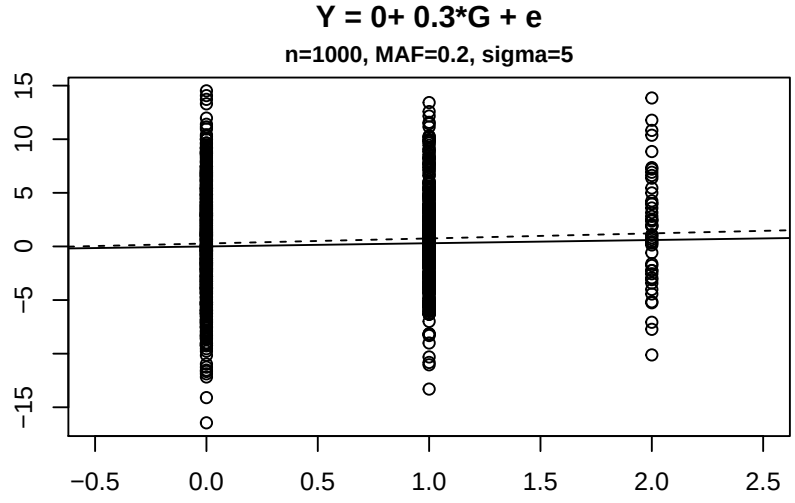
beta.0.hat=−0.05(0.04), beta.hat=0.34(0.05)



beta.0.hat=0.05(0.12), beta.hat=−0.24(0.2)



beta.0.hat=−0.02(0.03), beta.hat=0.44(0.1)



beta.0.hat=0.27(0.19), beta.hat=0.47(0.27)

What if Y is binary for a case-control study?

A continuous trait: $E(Y) = \beta_0 + \beta_j G_j$ (no E for notation simplicity)

A binary trait: $E(Y) = 1 \cdot \text{Prob}(Y = 1) + 0 \cdot \text{Prob}(Y = 0) = \text{Prob}(Y = 1)$

Instead of studying $E(Y) = \text{Prob}(Y = 1)$ directly, use a 'smart' transformation, $g(E(Y))$,

$$logit(E(Y)) = logit(\text{Prob}(Y = 1)) = log(\frac{\text{Prob}(Y = 1)}{1 - \text{Prob}(Y = 1)}) \in (-\infty, \infty)$$

Logistic regression, a generalized linear model (GLM),

$$g(E(Y)) = logit(E(Y)) = log(\frac{\text{Prob}(Y = 1)}{1 - \text{Prob}(Y = 1)}) = \beta_0 + \beta_j G_j$$

Interpretation: β **is the logOR** and

$$\text{Prob}(Y = 1) = \frac{exp(\beta_0 + \beta_j G_j)}{1 + exp(\beta_0 + \beta_j G_j)}$$

Binary trait simulation study (not discussed here)

NOT easy!

We can use **a liability/threshold model** to create cases and controls from a continuous outcome, similar to a population-based case-control study using e.g. the UK Biobank data.

Can we talk about $PRS_i = \sum_{j=1}^{J} \hat{\beta}_j G_{ij}$ now?
NOT yet: A deeper understanding of GWAS is needed!

Even determining J, the 'top' associated/ranked SNPs, is not that simple! Several complications:

- multiple hypothesis testing (mht; here)
- weak-moderate genetic effect size (low power; next)
- correlated tests (LD) ('power'?; a bit at the end)

  (many complex and interesting Qs, e.g. consider LD prior or post GWAS?)
- ...

mht: from $\alpha = 0.05$ to 5×10^{-8}, the genome-wide (GW) significance level

Dudbridge and Gusnanto (2008). *Genetic Epidemiology*. Estimation of significance thresholds for genomewide association scans.

$$g(E(Y)) = \beta_0 + \beta_j G_j; \quad H_0 : \beta_j = 0 \text{ for } j = 1, \ldots \approx 10^6 \text{ SNPs.}$$

- $\alpha = 0.05$**: many 'significant' SNPs per GWAS/family of tests**.
- If all SNPs are not associated, p-values are Unif(0,1) distributed.
- $\alpha = 5 \times 10^{-8}$**: family/GWAS-wise error rate (FWER) of 0.05**,

  Prob(at least one false positive SNP per GWAS) ≤ 0.05.

An illustrative simulation study: no SNPs associated

Using $\alpha = 0.05$ leads to 256 significant SNPs (all false positives) in this one single 'GWAS.null'! No significant findings at $\alpha = 0.05/5000$

```
set.seed(101)
nsample=1000; nsnp=5000 # less than 10^6 and no LD for now
G=matrix(-9,nrow = nsample,ncol = nsnp) # the genotype matrix
maf=runif(nsnp,min=0.05,max=0.5) # MAF randomly drawn from Unif(0,05,0.5) for simplicity
maf.hat=rep(-9,nsnp)
nsnp.true=0 # number of truly associated SNPs
beta.true=0 # no effect to study type 1 error
beta=c(rep(beta.true,nsnp.true),rep(0,(nsnp-nsnp.true))) # beta vector
betaG=rep(0,nsample) # the initial beta*G vector
for(j in 1:nsnp){ # using the loop function slows down the computation but adds clarity for teaching.
  nG=rmultinom(1,size=nsample,prob=c((1-maf[j])^2,2*maf[j]*(1-maf[j]), maf[j]^2))
  maf.hat[j]=(2*nG[3]+nG[2])/(2*nsample) # MAF estimated from the sample
  G[,j]=sample(c(rep(0,nG[1]),rep(1,nG[2]),rep(2,nG[3]))) # shuffle the G; no LD
  betaG=betaG+beta[j]*G[,j]
}
beta.0=0;sigma=1;e=rnorm(nsample,mean=0,sd=sigma)
Y=beta.0+betaG+e # the phenotype vector
sum.stat=matrix(-9,nrow=nsnp,ncol=7)
colnames(sum.stat)=c("MAF", "MAF.hat", "beta", "beta.hat", "se", "Z.value", "p.value")
for(j in 1:nsnp){
  fit=lm(Y~G[,j]); sum.stat[j,]=c(maf[j],maf.hat[j], beta[j], summary(fit)$coefficients[2,])
}
sum(sum.stat[,"p.value"]<=0.05) # many "significant" SNPs in this one single "GWAS"
```

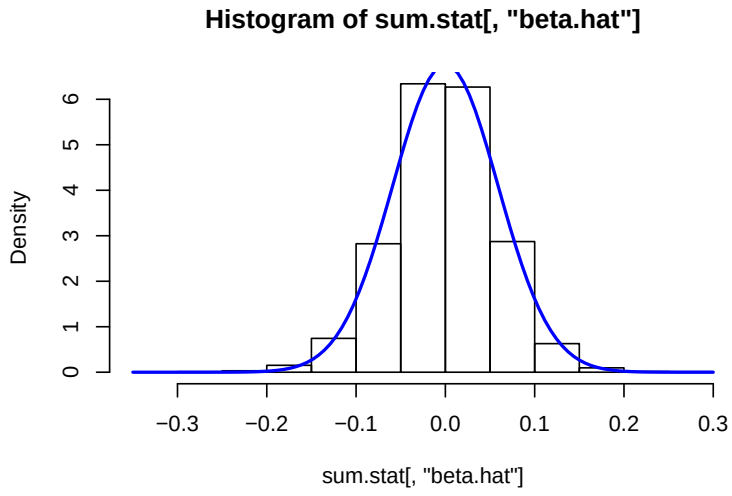
```
## [1] 256
```

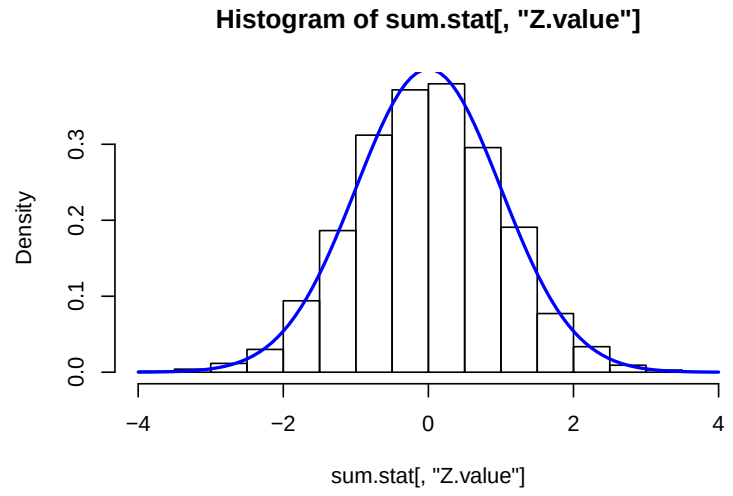
```
sum(sum.stat[,"p.value"]<=0.05/nsnp) # using the Bonferroni correction for FWER of 0.05
```

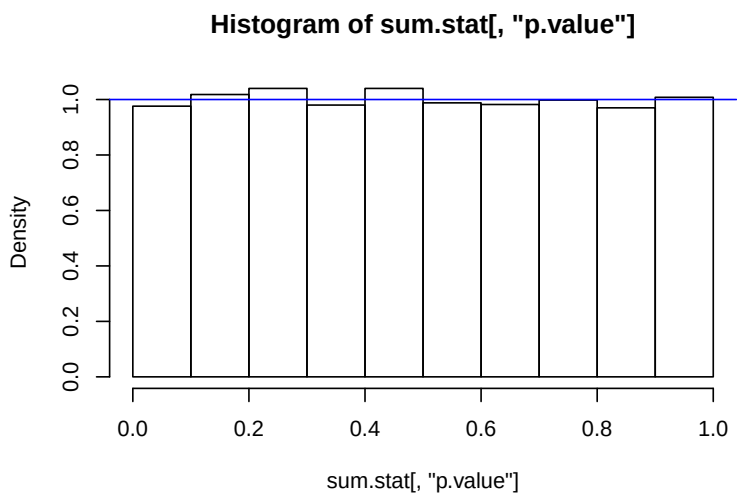
```
## [1] 0
```

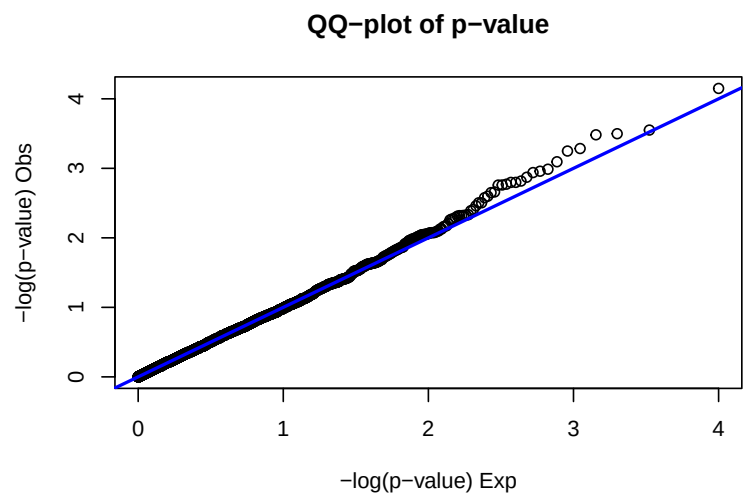
Pay attention to the spread of $\hat{\beta}_j$ histogram (top-left plot)

No association here, true $\beta_j = 0$ for all SNPs.

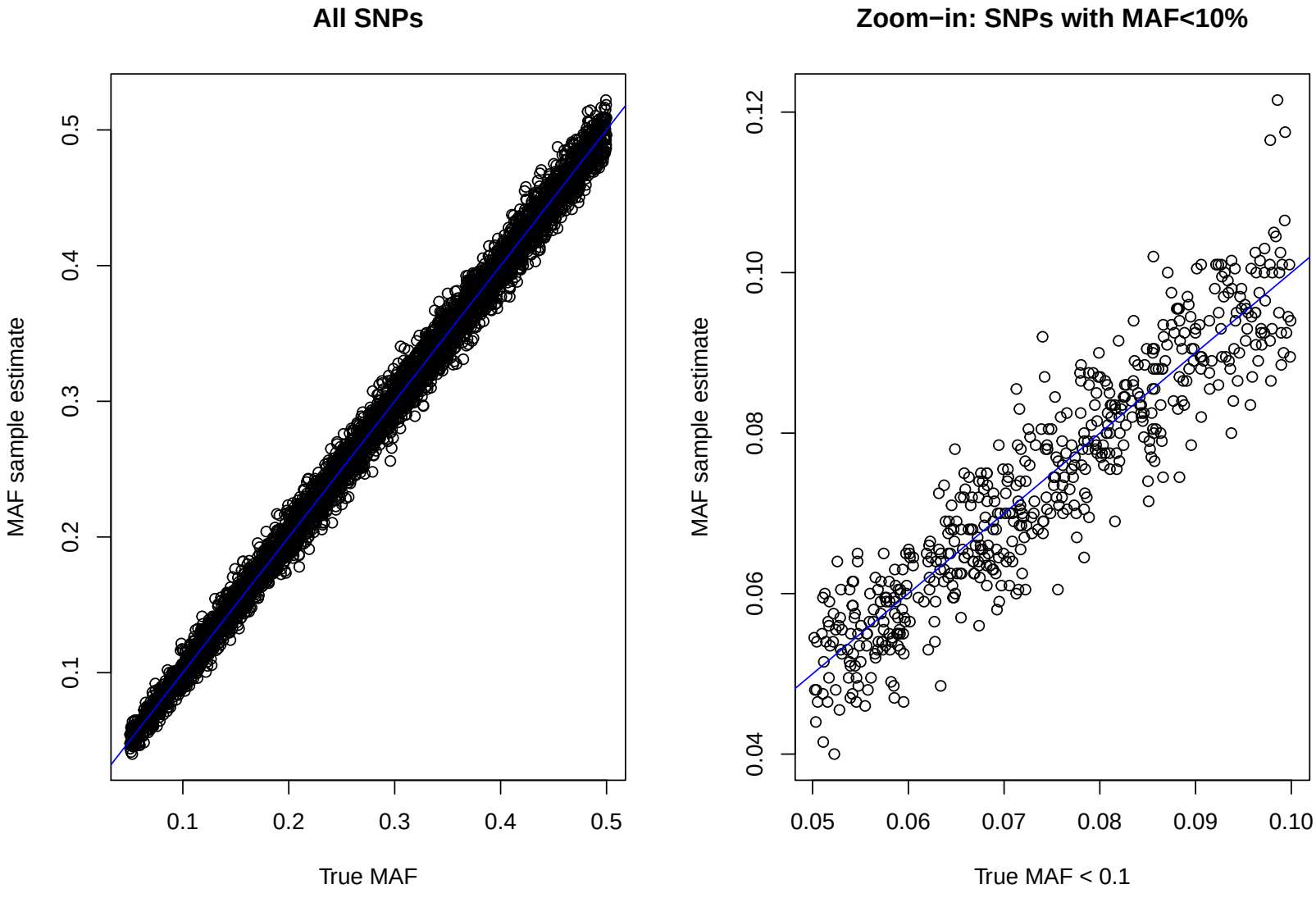








Also pay attention to the uncertainty in MAF estimates



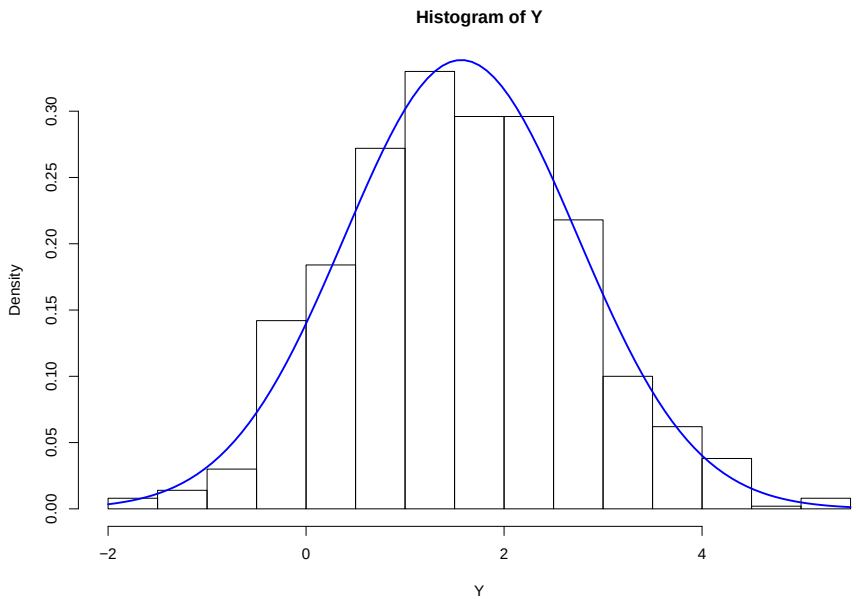
An illustrative 'polygenic' model simulation study

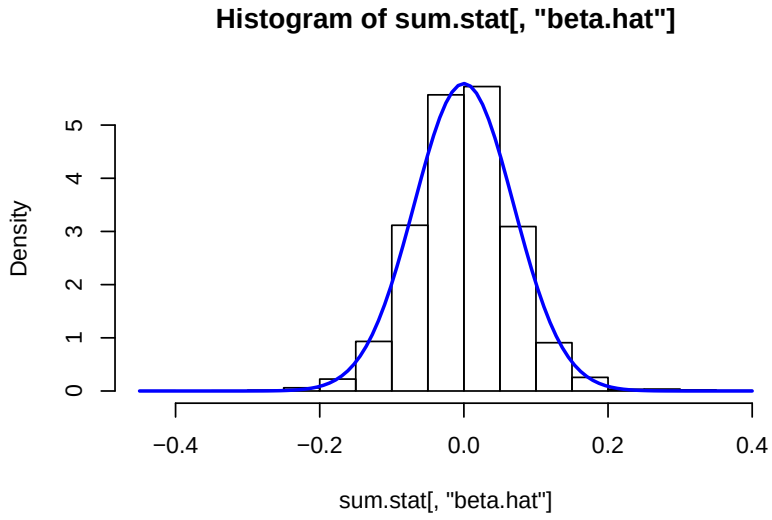
10 out 5000 indep. SNPs with **varying 'moderate-large' effects** are truly associated with Y (**all** $\beta = 0.3$ **but MAF vary**).

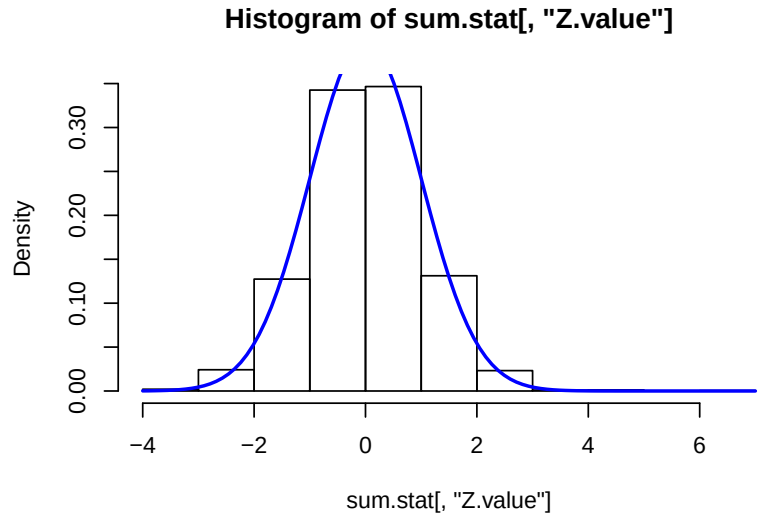
$$Y_i = \sum_{j=1}^{10} \beta_j G_{ij} + e, \text{ where } \beta_j = 0.3$$

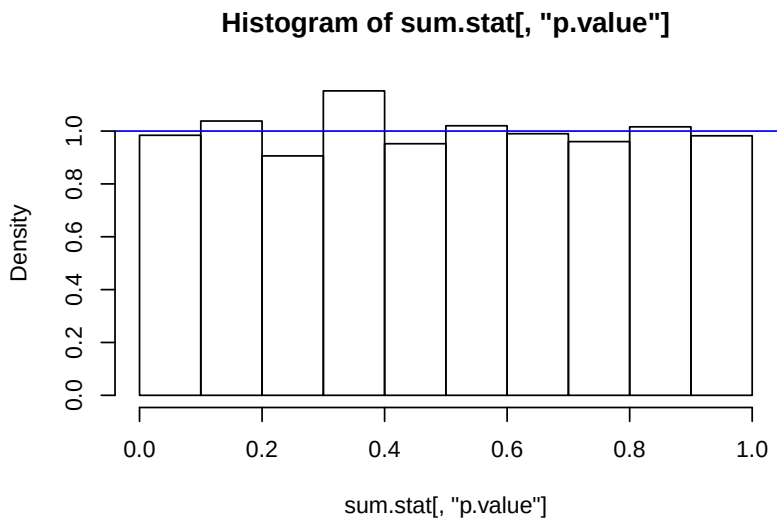
$$\text{MAF} \sim \text{ Unif}(0.05,0.5), \; e \sim N(0,1).$$

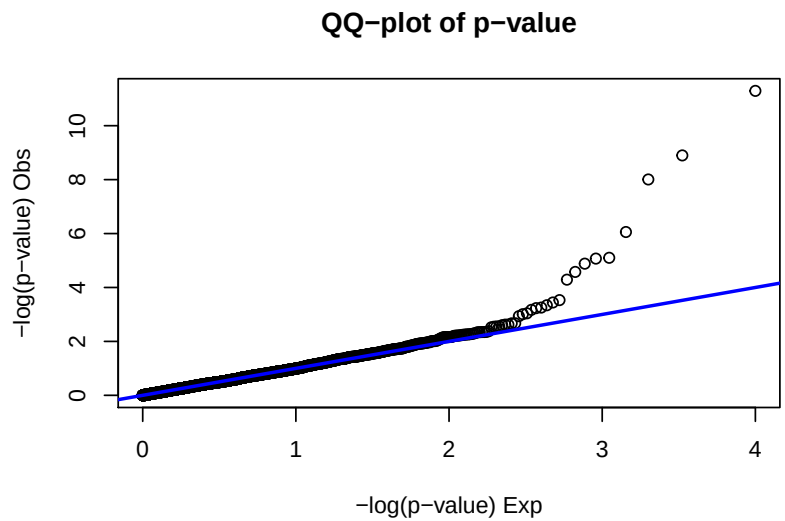
```
nsnp.true=10 # number of truly associated SNPs
beta.true=0.3 # "large" effect (also MAF, the error term, and the sample size)
```







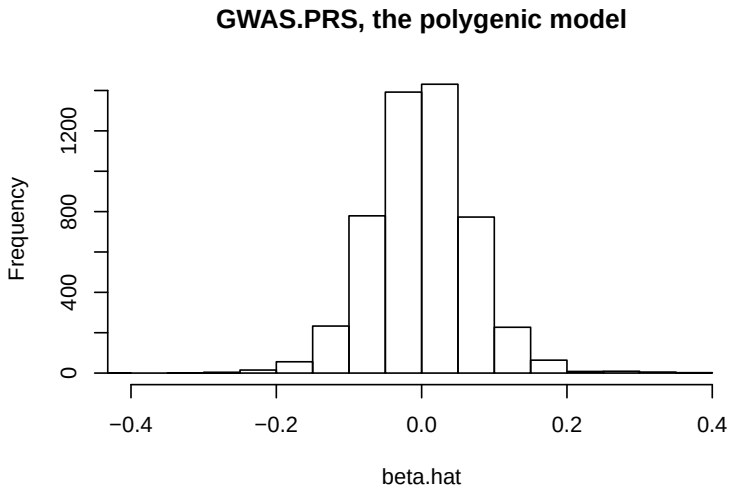


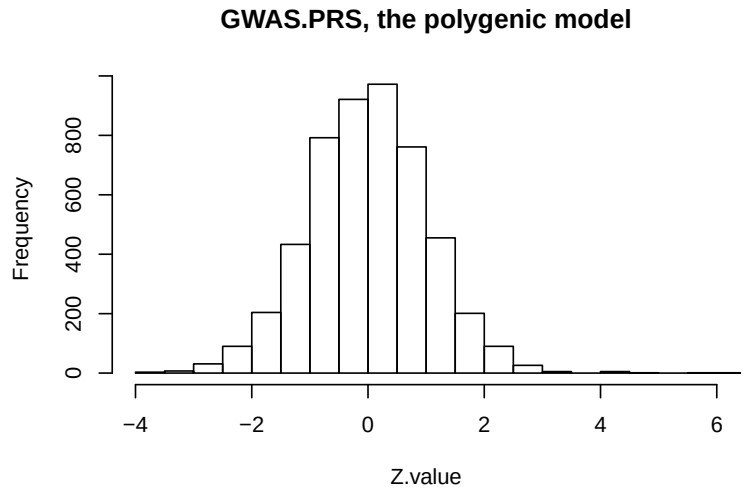


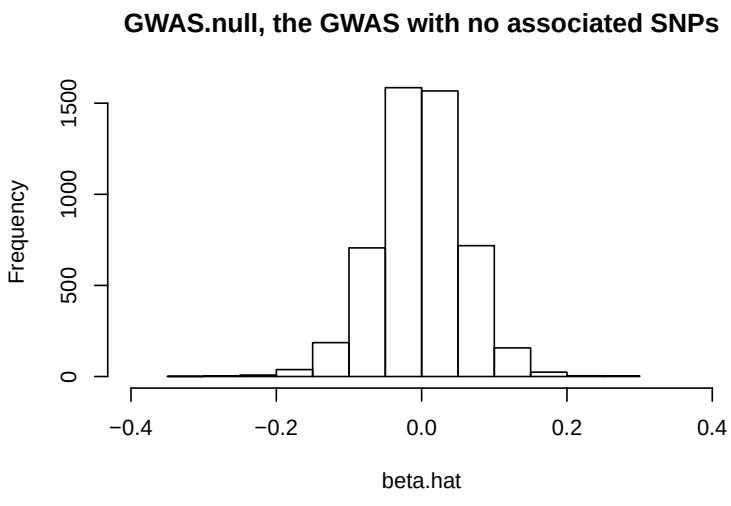
N.B. Histogram and QQ-plot carry different types of information!

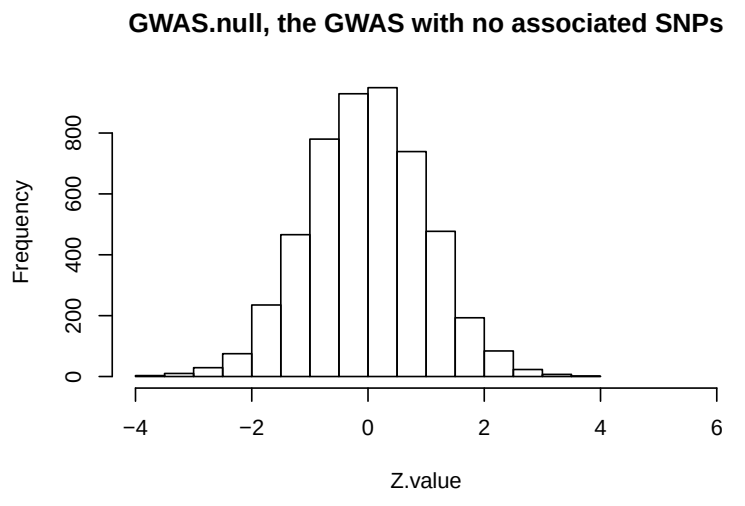
A closer look at the histograms of $\hat{\beta}$ and $Z = \hat{\beta}/SE$

Trouble ahead: similar between the GWAS.PRS (top) and GWAS.null (bottom)









GWAS-type of summary statistics

```
##              MAF MAF.hat beta    beta.hat         se     Z.value      p.value
##  [1,] 0.21748927  0.2215  0.3  0.29257288 0.06536792  4.47578705 8.489445e-06
##  [2,] 0.06972117  0.0610  0.3  0.33145758 0.10935747  3.03095507 2.500692e-03
##  [3,] 0.36935781  0.3780  0.3  0.23908858 0.05323916  4.49084039 7.922031e-06
##  [4,] 0.34596068  0.3480  0.3  0.38889542 0.05565755  6.98728935 5.116550e-12
##  [5,] 0.16243508  0.1695  0.3  0.30892955 0.07052329  4.38053229 1.308960e-05
##  [6,] 0.18502467  0.1995  0.3  0.37606430 0.06503910  5.78212649 9.859505e-09
##  [7,] 0.31318998  0.3375  0.3  0.33166110 0.05410586  6.12985587 1.264930e-09
##  [8,] 0.20006021  0.2020  0.3  0.28159164 0.06670313  4.22156565 2.647447e-05
##  [9,] 0.32990538  0.3360  0.3  0.23025579 0.05661344  4.06715744 5.134017e-05
## [10,] 0.29562285  0.2905  0.3  0.28906539 0.05841261  4.94868102 8.766086e-07
## [11,] 0.44590808  0.4445  0.0  0.09584075 0.05424572  1.76678916 7.756916e-02
## [12,] 0.36809363  0.3745  0.0 -0.02245388 0.05302784 -0.42343559 6.720687e-01
## [13,] 0.37938767  0.3750  0.0 -0.06366768 0.05424574 -1.17368994 2.407993e-01
## [14,] 0.46923549  0.4740  0.0  0.03095466 0.05222091  0.59276375 5.534736e-01
## [15,] 0.25480427  0.2485  0.0  0.05966600 0.06226877  0.95820104 3.381935e-01
## [16,] 0.31564388  0.3205  0.0 -0.03353920 0.05695716 -0.58884957 5.560954e-01
## [17,] 0.41919624  0.4345  0.0 -0.08589125 0.05307934 -1.61816710 1.059426e-01
## [18,] 0.15085332  0.1410  0.0  0.03167344 0.07632138  0.41500089 6.782304e-01
## [19,] 0.23525007  0.2515  0.0 -0.05445552 0.05876396 -0.92668222 3.543156e-01
## [20,] 0.06737475  0.0740  0.0 -0.10570983 0.10173334 -1.03908733 2.990158e-01
## [21,] 0.36532020  0.3595  0.0  0.06726877 0.05527611  1.21695922 2.239074e-01
## [22,] 0.48057686  0.4785  0.0  0.00804454 0.05286361  0.15217538 8.790794e-01
## [23,] 0.14600840  0.1400  0.0 -0.06318882 0.07458090 -0.84725206 3.970578e-01
## [24,] 0.34747768  0.3405  0.0 -0.00162502 0.05702567 -0.02849630 9.772720e-01
## [25,] 0.46549350  0.4590  0.0  0.07744989 0.05301480  1.46091065 1.443547e-01
## [26,] 0.40807389  0.4185  0.0  0.02135427 0.05247701  0.40692616 6.841495e-01
## [27,] 0.08204565  0.0820  0.0 -0.00372138 0.09588876 -0.03880935 9.690502e-01
## [28,] 0.22523350  0.2240  0.0 -0.09720419 0.06372866 -1.52528210 1.275056e-01
## [29,] 0.23290305  0.2455  0.0 -0.03611566 0.05999003 -0.60202767 5.472925e-01
## [30,] 0.34670979  0.3425  0.0  0.04700733 0.05368057  0.87568618 3.814114e-01
```

SNP1 output

```
##
## Call:
## lm(formula = Y ~ G[, 1])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3958 -0.8050  0.0109  0.7912  3.8070
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.43764    0.04691  30.647  < 2e-16 ***
## G[, 1]       0.29257    0.06537   4.476 8.49e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.167 on 998 degrees of freedom
## Multiple R-squared:  0.01968,    Adjusted R-squared:  0.0187
## F-statistic: 20.03 on 1 and 998 DF,  p-value: 8.489e-06
```

SNP2 output

```
##
## Call:
## lm(formula = Y ~ G[, 2])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2024 -0.8023 -0.0395  0.8466  3.7030
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.52681    0.03943  38.723   <2e-16 ***
## G[, 2]       0.33146    0.10936   3.031   0.0025 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.173 on 998 degrees of freedom
## Multiple R-squared:  0.009121,   Adjusted R-squared:  0.008128
## F-statistic: 9.187 on 1 and 998 DF,  p-value: 0.002501
```

IF we knew which set of SNPs to include (getting into the PRS direction but not yet)

```
# not realistic: no GWAS needed if we already know which SNPs are associated!
summary(lm(Y~G[,1]+G[,2]+G[,3]+G[,4]+G[,5]+G[,6]+G[,7]+G[,8]+G[,9]+G[,10]))
```

```
##
## Call:
## lm(formula = Y ~ G[, 1] + G[, 2] + G[, 3] + G[, 4] + G[, 5] +
##     G[, 6] + G[, 7] + G[, 8] + G[, 9] + G[, 10])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2616 -0.7070 -0.0004  0.6983  3.3557
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.03884    0.09679  -0.401  0.68830
## G[, 1]       0.30033    0.05783   5.193 2.51e-07 ***
## G[, 2]       0.29903    0.09607   3.112  0.00191 **
## G[, 3]       0.31183    0.04731   6.592 7.06e-11 ***
## G[, 4]       0.35997    0.05006   7.190 1.27e-12 ***
## G[, 5]       0.32629    0.06260   5.212 2.27e-07 ***
## G[, 6]       0.38079    0.05782   6.586 7.33e-11 ***
## G[, 7]       0.31377    0.04840   6.482 1.42e-10 ***
## G[, 8]       0.32870    0.05901   5.570 3.28e-08 ***
## G[, 9]       0.27331    0.05002   5.464 5.90e-08 ***
## G[, 10]      0.27448    0.05183   5.296 1.46e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.027 on 989 degrees of freedom
## Multiple R-squared:  0.2471, Adjusted R-squared:  0.2395
## F-statistic: 32.46 on 10 and 989 DF,  p-value: < 2.2e-16
```

What kind of model (heritability, h^2) did we simulate?

$$Y = \sum_{j=1}^{10} \beta_j G_j + e, \text{ where } \beta_j = 0.3, \ e \sim N(0, \sigma^2 = 1),$$

The MAF of the 10 truly associated SNPs, ranging from 0.07 to 0.37:

```
round(maf[1:nsnp.true],2)
```

```
##  [1] 0.22 0.07 0.37 0.35 0.16 0.19 0.31 0.20 0.33 0.30
```

True (not estimated) heritability:

$$h^2 = \frac{V_G}{V_G + V_e} = \frac{0.321}{0.321 + 1} = 24.3\%$$

```
V.G=sum(beta[1:nsnp.true]^2*(2*maf[1:nsnp.true]*(1-maf[1:nsnp.true])))
V.e=sigma^2
h2=V.G/(V.G+V.e)
round(c(V.G,V.e,h2),3)
```

```
## [1] 0.321 1.000 0.243
```

Analytical details for h^2 of this **simple** model

(linear, fixed-effect, additive, no LD, no interaction)

$$Y = \sum_{j=1} \beta_j G_j + e, \text{ where } e \sim N(0, \sigma^2).$$

$$V_Y = Var(Y) = \sum_j \beta_j^2 Var(G_j) + \sigma^2 = V_G + V_e,$$

- G_j: p_j as MAF for A
- coded additively: $0 = aa$, $1 = Aa$ and $2 = AA$
- genotype frequency under HWE: $(1-p_j)^2$, $2p_j(1-p_j)$ and p_j^2
- $E(G_j) = 2p_j$; $Var(G_j) = 2p_j(1-p_j)$

$$\text{(narrow) } h^2 = \frac{V_G}{V_G + V_e} = \frac{\sum_j \beta_j^2 Var(G_j)}{Var(Y)} = \frac{\sum_j \beta_j^2 2p_j(1-p_j)}{\sum_j \beta_j^2 2p_j(1-p_j) + \sigma^2}.$$

Heritability of GWAS 'loci', h_j^2 contributed by each individual, independent SNP j in our case

$$\text{(narrow) } h_j^2 = \frac{\beta_j^2 2p_j(1-p_j)}{\sum_j \beta_j^2 2p_j(1-p_j) + \sigma^2}.$$

$\beta_j = 0.3$ **for all 10 causal SNPs but MAFs differ:**

```
##  [1] 0.217 0.070 0.369 0.346 0.162 0.185 0.313 0.200 0.330 0.296
```

Thus, (true not estimated) SNP h_j^2's differ:

```
##  [1] 0.025 0.010 0.034 0.033 0.020 0.022 0.032 0.024 0.033 0.031
```

Worth repeating: What is the effect size of a SNP?

All 10 SNPs have $\beta = 0.3$, but their (true not estimated) h^2 contributions vary

from 1% (MAF=0.07, SNP2)

to 3.4% (MAF=0.37, SNP3)

$\sum_j \beta_j^2 2p_j(1-p_j)$

Effect interpretation depends on MAF (and also σ^2).

n comes in later when we try to find these SNPs using data.

In practice, β_j must be estimated and

Large n is then critical!

MAF p_j and σ^2 also need to be estimated.

Connection with Explained Variatation (EV) and R^2 from regression

(linear, fixed-effect, additive, no LD, no interaction)

$$Y = \sum_{j=1} \beta_j G_j + e, \text{ where } e \sim N(0, \sigma^2).$$

$$EV = \frac{\text{variation of Y explained by G}}{\text{total variation of Y}} = \frac{Var(E(Y|G))}{Var(Y)}$$
$$= \frac{\sum_j \beta_j^2 2p_j(1-p_j)}{\sum_j \beta_j^2 2p_j(1-p_j) + \sigma^2} = h^2.$$

$$R^2 = \frac{SS_{explained}}{SS_{total}} = 1 - \frac{SS_{residual}}{SS_{total}}$$
$$= 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \approx h^2$$

Recall: multi-SNP regression **IF** we knew the true model

```
##
## Call:
## lm(formula = Y ~ G[, 1] + G[, 2] + G[, 3] + G[, 4] + G[, 5] +
##     G[, 6] + G[, 7] + G[, 8] + G[, 9] + G[, 10])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2616 -0.7070 -0.0004  0.6983  3.3557
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.03884    0.09679  -0.401  0.68830
## G[, 1]       0.30033    0.05783   5.193 2.51e-07 ***
## G[, 2]       0.29903    0.09607   3.112  0.00191 **
## G[, 3]       0.31183    0.04731   6.592 7.06e-11 ***
## G[, 4]       0.35997    0.05006   7.190 1.27e-12 ***
## G[, 5]       0.32629    0.06260   5.212 2.27e-07 ***
## G[, 6]       0.38079    0.05782   6.586 7.33e-11 ***
## G[, 7]       0.31377    0.04840   6.482 1.42e-10 ***
## G[, 8]       0.32870    0.05901   5.570 3.28e-08 ***
## G[, 9]       0.27331    0.05002   5.464 5.90e-08 ***
## G[, 10]      0.27448    0.05183   5.296 1.46e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.027 on 989 degrees of freedom
## Multiple R-squared:  0.2471, Adjusted R-squared:  0.2395
## F-statistic: 32.46 on 10 and 989 DF,  p-value: < 2.2e-16
```

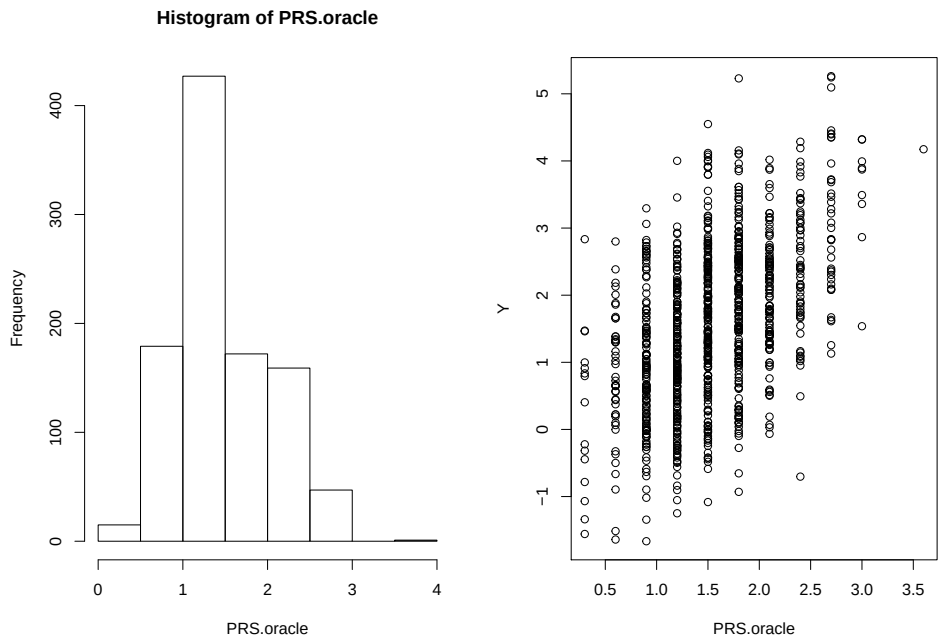
From multi-SNP to one-super-SNP (PRS) association!

IF we knew the true model, we can construct PRS_{oracle}

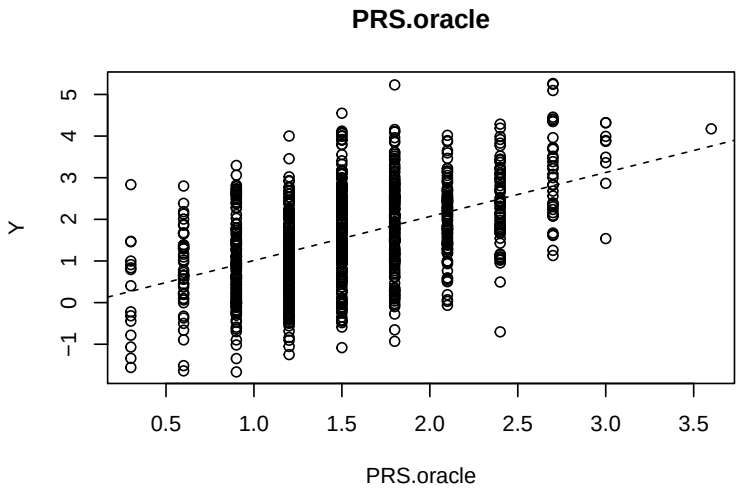
Not realistic: only to demonstrate the value of PRS **conceptually**.

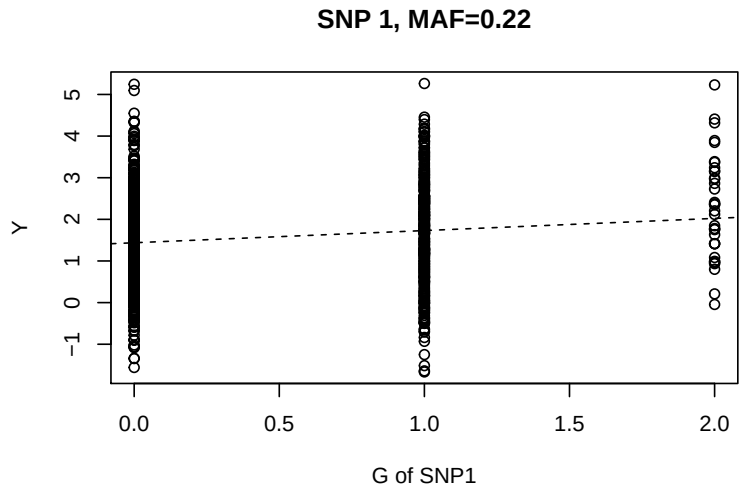
$$PRS_{i,oracle} = \sum_{j=1}^{J=10} \beta_j G_{ij} + e, \text{ where } \beta_j = 0.3$$

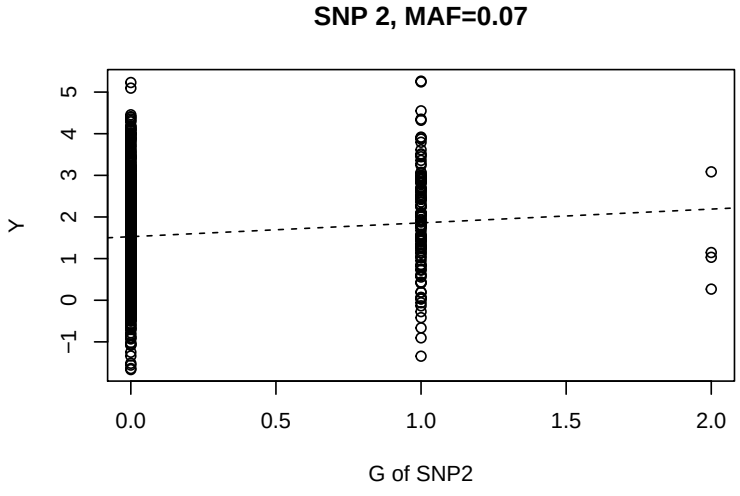
```
PRS.oracle=rep(0,nsample)  # the PRS vector
for (i in 1:nsample) # for each individual i
  for (j in 1:nsnp.true) # sum over the J selected SNPs
    PRS.oracle[i] = PRS.oracle[i]+beta[j]*G[i,j]
par(mfrow=c(1,2))
hist(PRS.oracle) # not quite normal as J=10 here
plot(PRS.oracle,Y) # much more predictive than individual SNPs
```

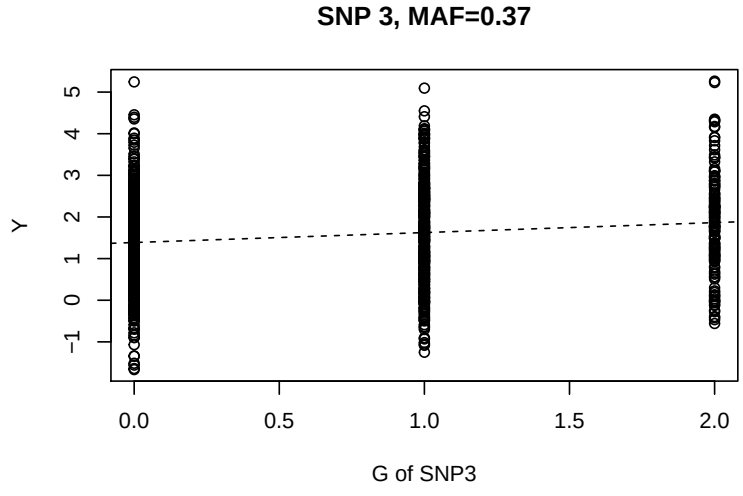


(Good) PRS is more significantly associated with the trait than one single SNP









```
summary(lm(Y~PRS.oracle))
```

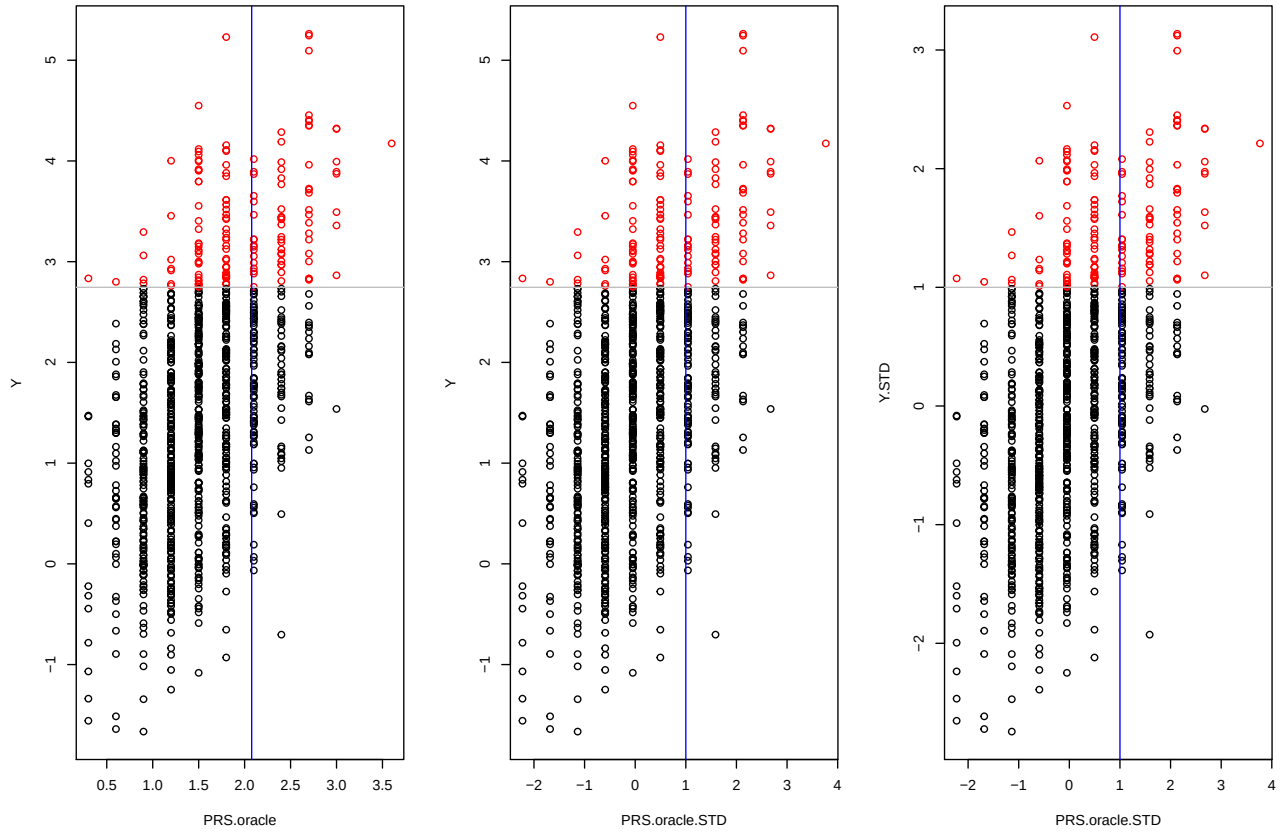
```
##
## Call:
## lm(formula = Y ~ PRS.oracle)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1940 -0.7168 -0.0158  0.7102  3.3731
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.04598    0.09549  -0.482     0.63
## PRS.oracle   1.05710    0.05886  17.959   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.025 on 998 degrees of freedom
## Multiple R-squared:  0.2442, Adjusted R-squared:  0.2435
## F-statistic: 322.5 on 1 and 998 DF,  p-value: < 2.2e-16
```

From PRS-based association to PRS-based prediction!

Standardization (STD) and a liability/threashold model

```
Y.STD=(Y-mean(Y))/sqrt(var(Y))
PRS.oracle.STD=(PRS.oracle-mean(PRS.oracle))/sqrt(var(PRS.oracle))
case.index=which(Y.STD>1);control.index=which(Y.STD<=1) # 1 is a subjective choice
c(length(Y[case.index]),length(Y[control.index])) # numbers of cases and controls
```

```
## [1] 149 851
```



higher PRS_{oracle} $\Longrightarrow$ higher risk (proportionally more case)

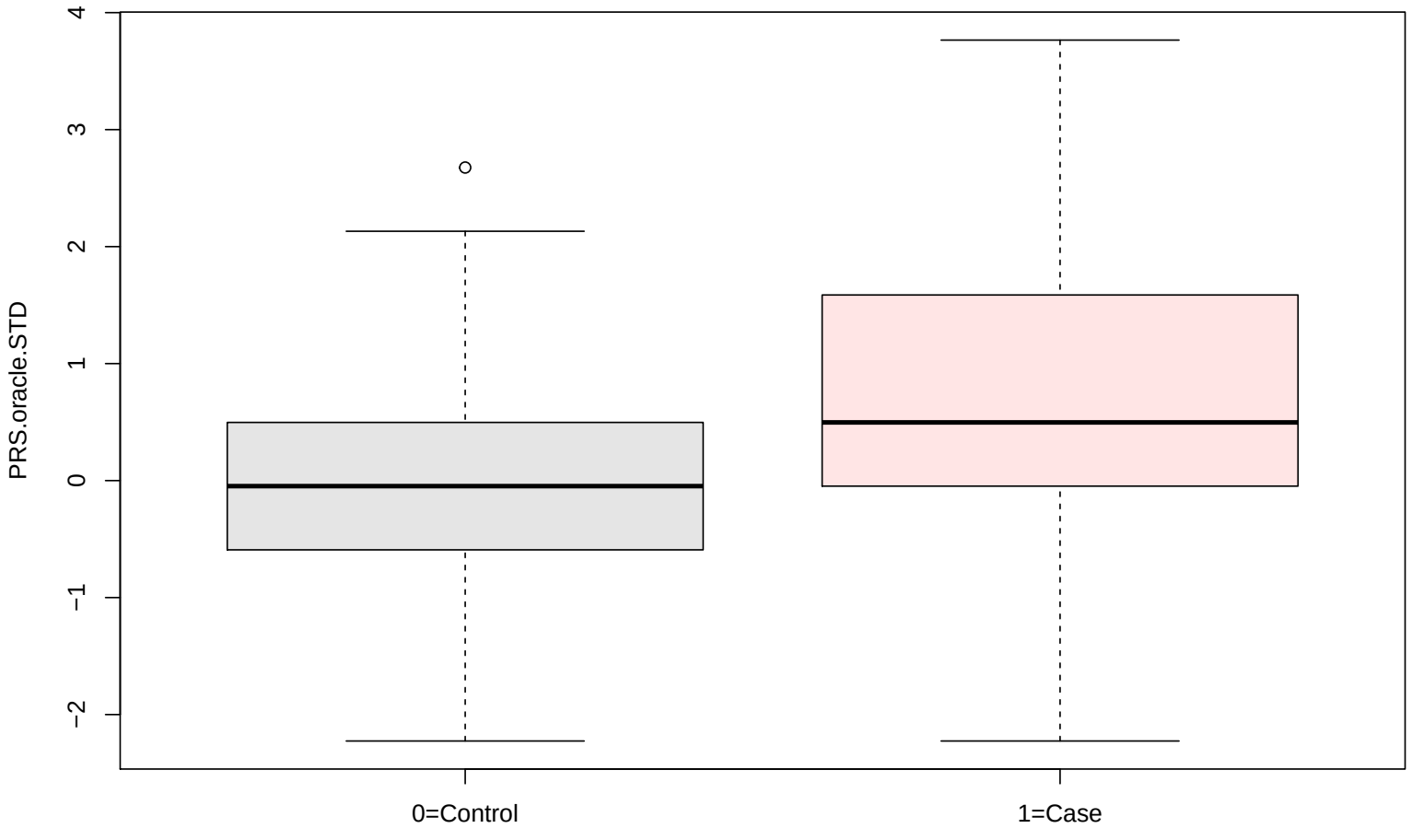
Quiz

Standardization (STD) is often done in practice and should not change interpretation.

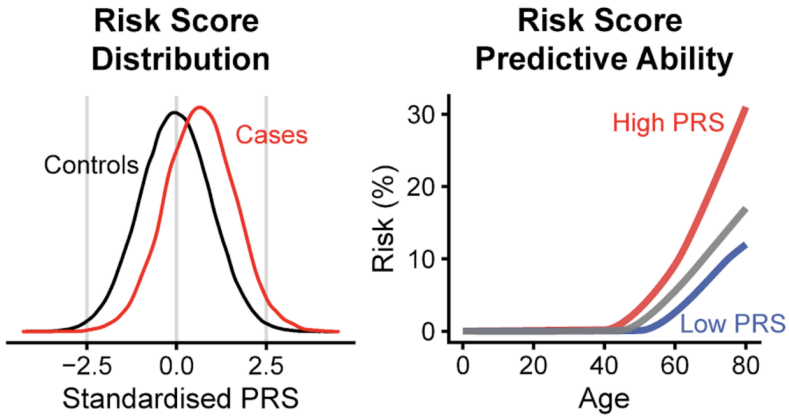
BUT, what are the potential pitfalls of STD?

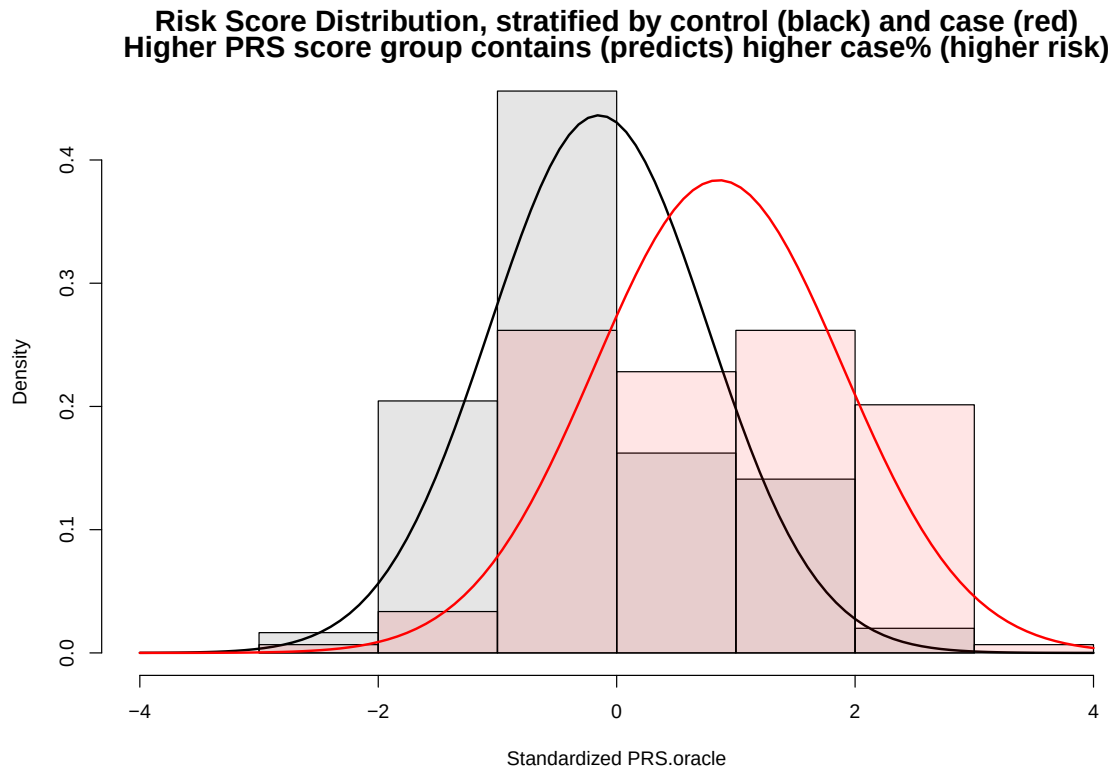
Different perspective but the same idea: PRS_{oracle}'s of cases tend to be higher than PRS_{orcale}'s of controls

```
Y.cc=rep("0=Control",nsample); Y.cc[case.index]="1=Case"
boxplot(PRS.oracle.STD~Y.cc,main="",xlab="",col=c(rgb(0,0,0,0.1),rgb(1,0,0,0.1)))
```



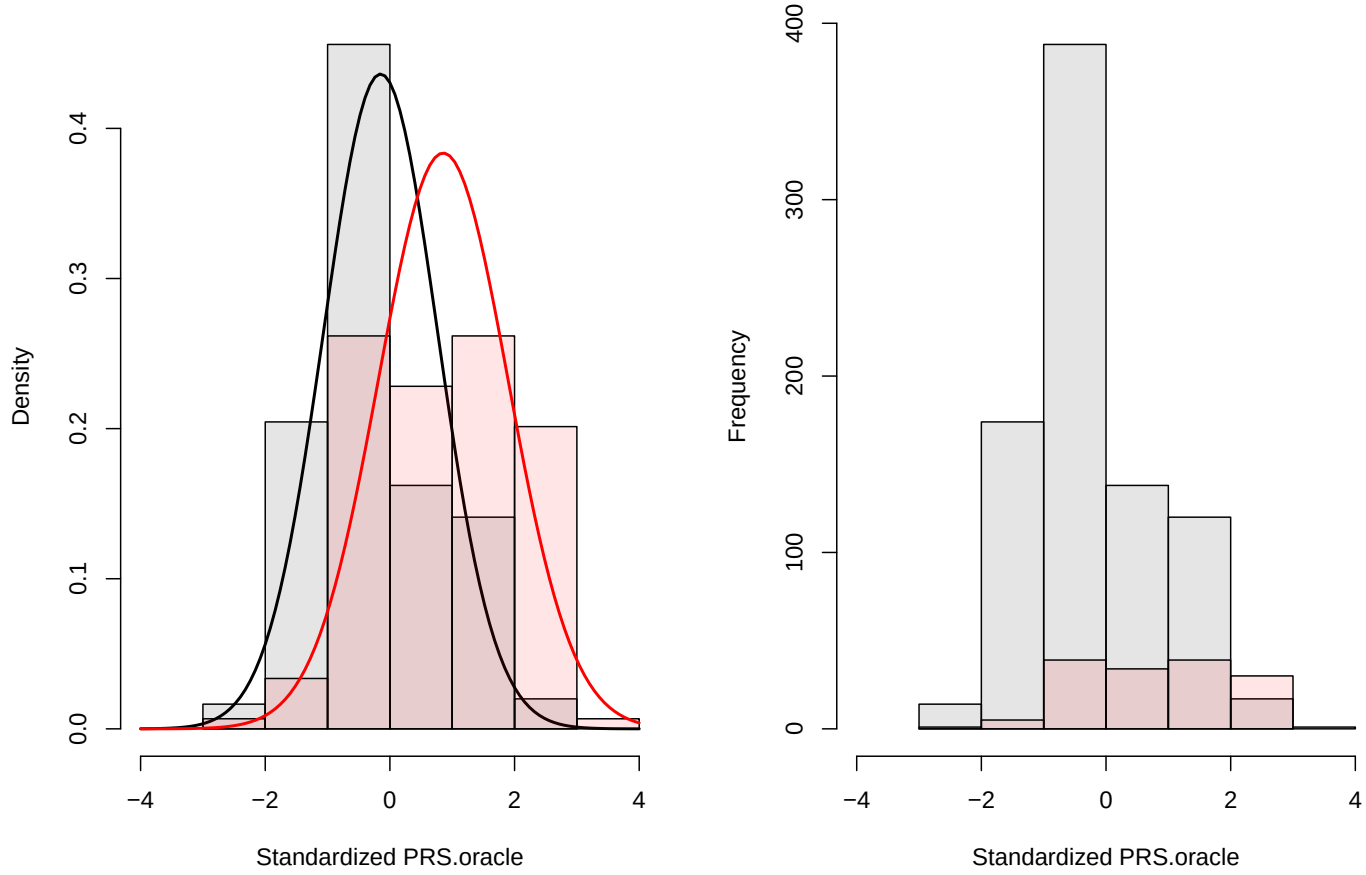
Recall and mimic the illustrative plots (2 in 1)





Quiz:

The standardized PRS.oracle value of an individual is 2.5. What is the **probability** of this individual having the disease/condition? (Hint in the two histrograms below and **relative risk** $\neq$ **absolute risk!**)

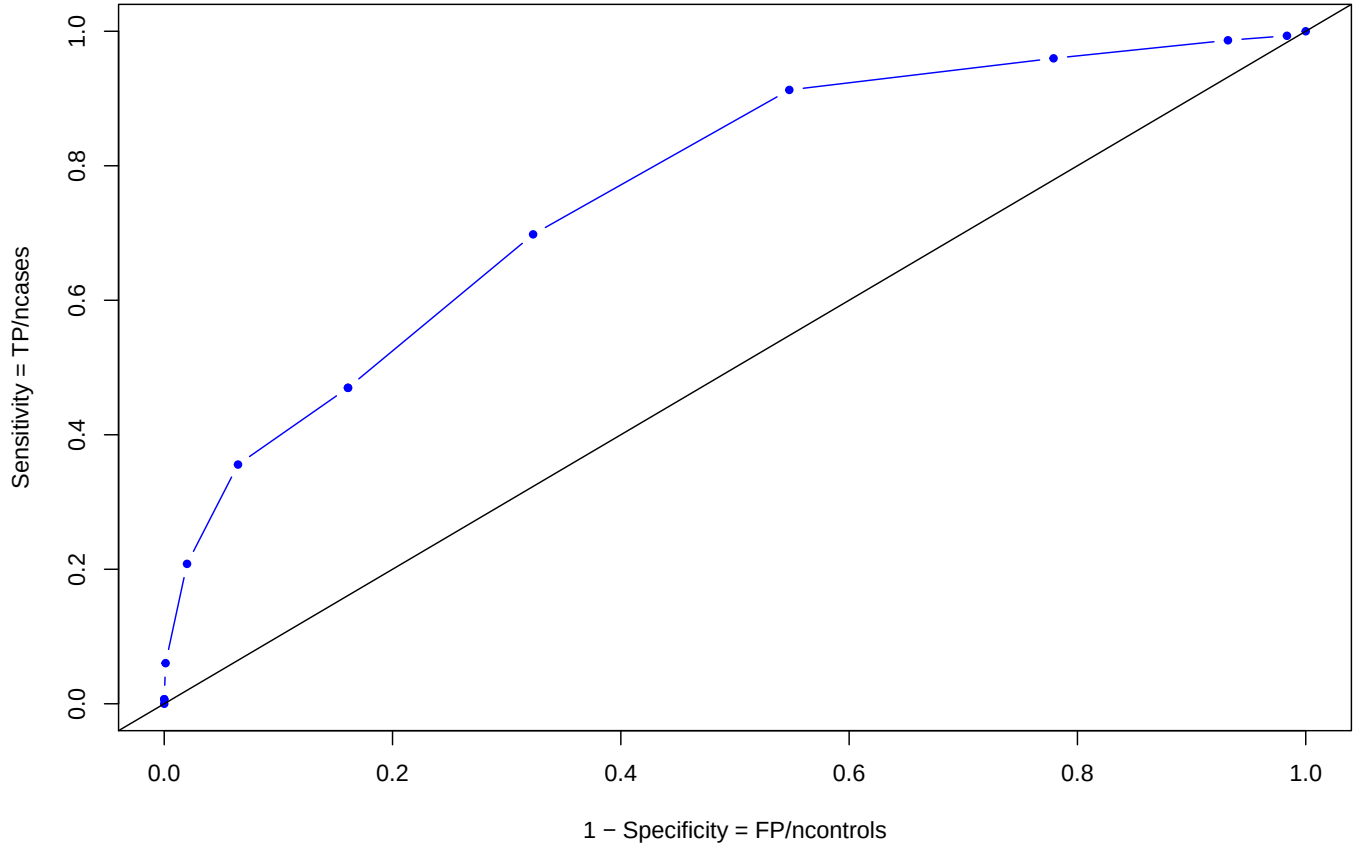


Another related quiz:

- cases: individuals with the disease/condition
- controls: individuals without the disease/condition
- a test or a decision rule, say Covid-19 testing or PRS-based prediction (e.g. standardized PRS >3 predicting case)
- Sensitivity = e.g. 90%
  $$\text{Sensitivity} = \Pr(\text{positive test result}|\text{case})$$
- Specificity = e.g. 90%
  $$\text{Specificity} = \Pr(\text{negative test result}|\text{control})$$

Is it possible that $\Pr(\text{case}|\text{PRS}>3) < 50\%$?
(hint: which information is missing from the above?)

Towards ROC (receiver operating characteristic) curve and AUC (area under the curve), using our simulated data and PRS.oracle



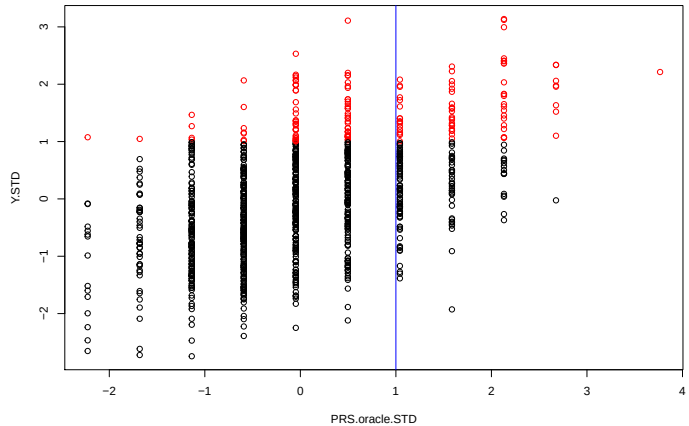

```
## [1] "AUC of ROC.oracle=" "0.763"
```

Undertanding each point on the ROC curve

```
plot(PRS.oracle.STD,Y.STD,col=color.index)
ncase=sum(Y.cc=="1=Case") # total number of cases
ncontrol=sum(Y.cc=="0=Control") # total number of controls
c(ncase,ncontrol)
```

```
## [1] 149 851
```

```
PRS.threshold=1;abline(v=PRS.threshold,col="blue")  # threshold used to call a sample positive/case
```




```
P=sum(PRS.oracle.STD>PRS.threshold) # number of Positives at this threshold
TP=sum(Y.cc[PRS.oracle.STD>PRS.threshold]=="1=Case") # True Positives
FP=sum(Y.cc[PRS.oracle.STD>PRS.threshold]=="0=Control") # False Positives
c(P,TP,FP)
```

```
## [1] 207  70 137
```

```
sensitivity=TP/ncase # sensitivity
specificity.1=FP/ncontrol # 1-specificity
c(sensitivity,specificity.1) # ONE point on the ROC curve: (y=sensitivity=0.47,x=1-specificity=0.16)
```

```
## [1] 0.4697987 0.1609871
```

A few more (sensitivity vs. 1-specificity) points for ROC

```
# increase the threshold: both sensitivity and 1-specificity decrease
PRS.threshold= -1.5
```

```
## [1] 940 147 793
```

```
## [1] 0.9865772 0.9318449
```

```
PRS.threshold=0
```

```
## [1] 379 104 275
```

```
## [1] 0.6979866 0.3231492
```

```
PRS.threshold=1.5
```

```
## [1] 108  53  55
```

```
## [1] 0.35570470 0.06462985
```

```
PRS.threshold=2
```

```
## [1] 48 31 17
```

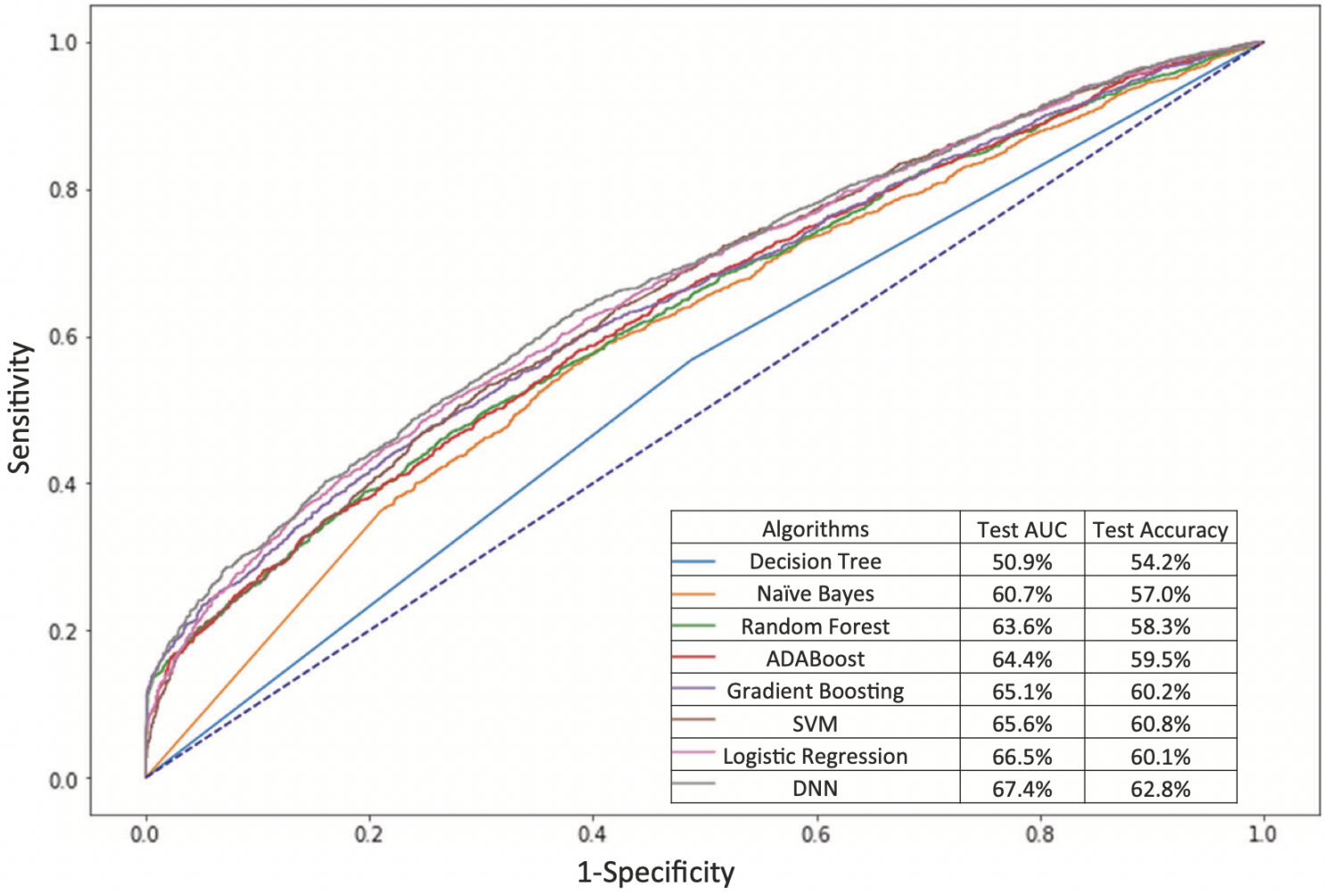
```
## [1] 0.2080537 0.0199765
```

```
PRS.threshold=2.5  # Q: variability of these estimates?
```

```
## [1] 10  9  1
```

```
## [1] 0.060402685 0.001175088
```

Real-life ROC curves, e.g.



Algorithms	Test AUC	Test Accuracy
Decision Tree	50.9%	54.2%
Naïve Bayes	60.7%	57.0%
Random Forest	63.6%	58.3%
ADABoost	64.4%	59.5%
Gradient Boosting	65.1%	60.2%
SVM	65.6%	60.8%
Logistic Regression	66.5%	60.1%
DNN	67.4%	62.8%

Figure 3 of Badre et al. (2021). *Journal of Human Genetics*. Deep neural network improves the estimation of polygenic risk scores for breast cancer.

N.B. The classical logistic regression is competitive!

Because,

$$\mathbf{PRS_{i,oracle}} = \sum_{j=1}^{J=10} \beta_j (= 0.3) G_{ij} \textbf{ is NOT } \mathbf{PRS_{i,parctice}}!$$

- J is unknown, to be determined
- β_j is unknown, to be estimated
- G_{ij} **cannot be directly from the same data used to infer** J **and** β_j.

  Otherwise: over-fitting/double-dipping/data-dredging/p-hacking/selection-bias!
- Not to mention LD and other considerations in real data settings.

What's next: HOW to construct $\mathbf{PRS_{practice}}$ and do it CORRECTLY!

Recap the goal of this lecture: a **deeper** understanding of

- the multiple hypothesis testing issue inherent in GWAS
- the (high) variability inherent in $\hat{\beta}$, the β estimates
- heritability h^2 as a function of both β and MAF (and σ^2)
- the 'genetic effect size' of a SNP $= \beta^2 \cdot \text{MAF} \cdot (1 - \text{MAF})$
- a conceptual PRS construction based on the ground truth, PRS.oracle
- DIY ROC plotting and AUC calculation for a PRS-based prediction

What's next: How to construct $PRS_{practice}$ and do it correctly and compare results using different α level.