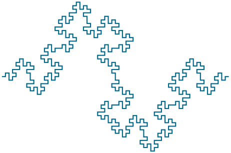
# High-Throughput Sequencing Course

## Statistical Inference: Sources of Variability

Biostatistics and Bioinformatics

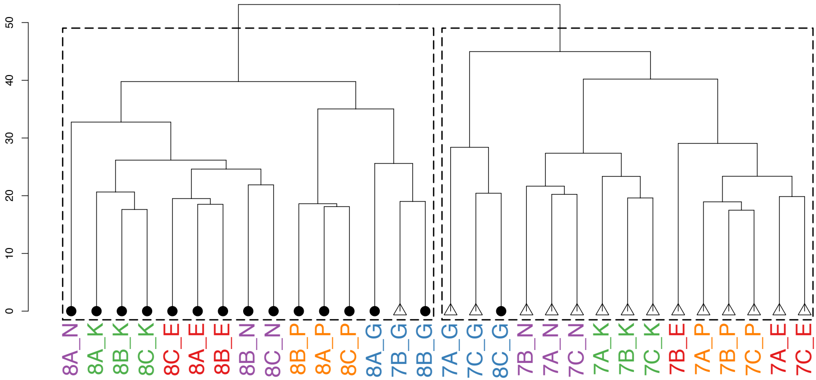


Summer 2017

Duke University
School of Medicine

DUKE BIOSTATISTICS
& BIOINFORMATICS

# Class Discovery



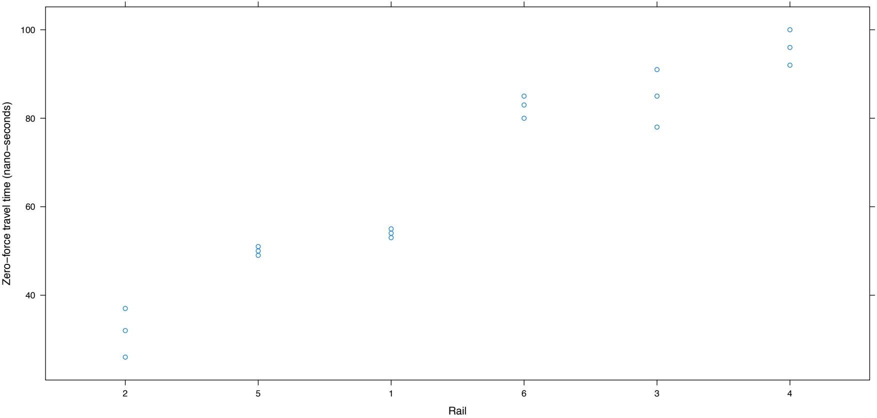
# Intra- and Inter-subject Variability

- In most experiments, including RNA-Seq, the variability may not be exclusively due to measurement error
- Another source could be due to repeated measurements
- or sampling from strains or cell lines
- or due to batch effects (e.g., team effect)
- We will motivate these ideas using a classical toy example
- We will illustrate the caveats of properly accounting for these two sources of variability through two simulation studies

# Rails Data

- Observation adjusted travel time for ultrasonic head-waves in the rail (nanoseconds).
- Data set: 6 rails; the travel time is sampled three times per rail
- Eighteen measurements
- Six Experimental Units
- Implicit assumption: The six rails are randomly selected from a *large* pool of rails
- What is of interest is neither the batch or any of these 6 rails (specifically)
- What is of interest is the population (the huge pool)

# Rail Data



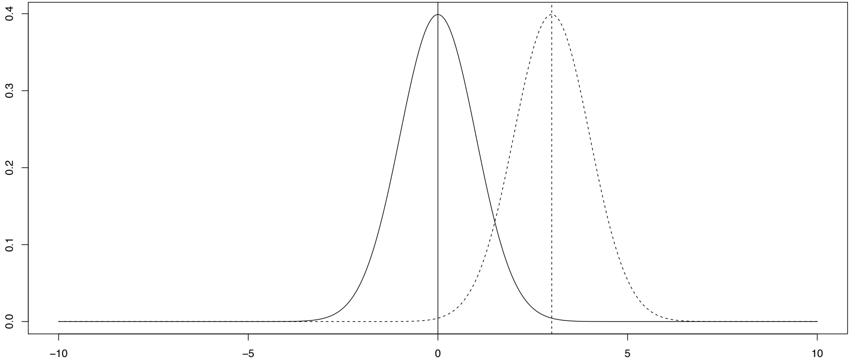
# Rail Data: Model Formulation

- $\mu$ denotes the *true* travel time
- $\mu$ is an unknown fixed quantity
- $Y_i$ denotes the *observed* travel time (for observation $i = 1, \ldots, 18$)
- In absence of noise, true value $\mu$ is observed
- In other words, $Y_i = \mu$ for $i = 1, \ldots, 18$

# Important Fact about Normal Distribution

- Consider a normal distribution with mean 0 and standard deviation $\sigma$
- If the data are shifted by a constant $\mu$, then
  1. resulting distribution remains normal
  2. The mean of the new distribution is $\mu + 0 = \mu$
  3. Its standard deviation remains unchanged
- The last two (but not first) property are true for any distribution

# Shift Normal Distribution



# Rail Data: Simple Model

- What is observed is a distorted version of $\mu$

$$Y_i = \mu + \epsilon_i$$

- Notes:
  - $Y_i$ is observable
  - $\epsilon_i$ is *not* observable
  - $\mu$ is an unknown parameter
- The variability observed here is exclusively attributed to the measurement error $\epsilon_i$

# Linear Model

```
summary(lm(travel ~ 1, data = Rail))
```

```
##
## Call:
## lm(formula = travel ~ 1, data = Rail)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -40.50 -16.25   0.00  18.50  33.50
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   66.500      5.573   11.93  1.1e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.65 on 17 degrees of freedom
```

# Rail Data: Account for Two Source of Variability

- What is observed is a distorted version of $\mu$
- It is distorted by a ra
- $Y_{ij}$: Index the rail by $i = 1, \ldots, 6$ and the replicate by $j = 1, 2, 3$
- $Y_{23}$: The obeservation for the third replicate for rail 2
- Model

$$Y_{ij} = \mu + b_i + \epsilon_{ij}$$

- Notes:
  - $Y_{ij}$ is observable
  - $b_i$ is *not* observable
  - $\epsilon_{ij}$ is *not* observable
  - $\mu$ is an unknown parameter

# Linear Mixed Effects Model

```r
lme(travel ~ 1, random = ~1 | Rail, data = Rail)
```

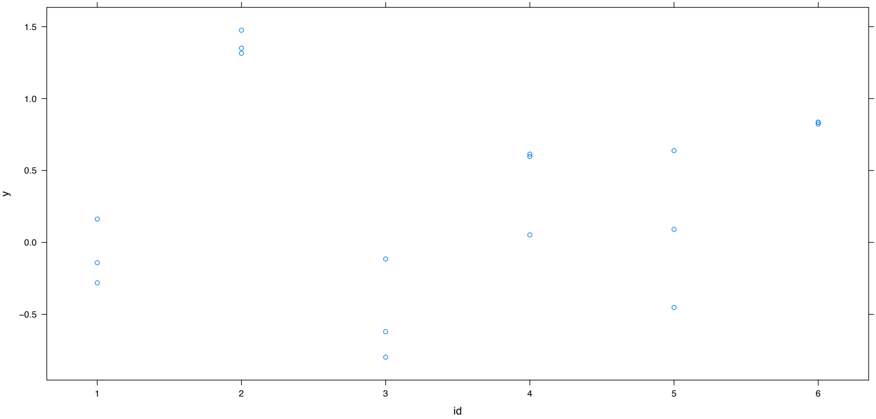
```
## Linear mixed-effects model fit by REML
##   Data: Rail
##   Log-restricted-likelihood: -61.0885
##   Fixed: travel ~ 1
## (Intercept)
##        66.5
##
## Random effects:
##  Formula: ~1 | Rail
##         (Intercept) Residual
## StdDev:    24.80547 4.020779
##
## Number of Observations: 18
## Number of Groups: 6
```

# Is the Mixed Model Adequate?

- Assumptions:
  - $b_i$ is normally distributed $N[0, \sigma_b^2]$
  - $\sigma_b^2$ does *not* depend on $i$ (homoscedastic)
  - $\epsilon_{ij}$ is normally distributed $N[0, \sigma_e^2]$
  - $\sigma_e^2$ does *not* depend on $i$ or $j$ (homoscedastic)
  - Error model is additive (could be multiplicative)

# Example 1: Setup

- What are the ramifications for ignoring the clustering?
- We will sample 6 experimental units each with three replicates
- $\mu = 0, \sigma_e = 0.25, \sigma_b = 0.5$



# Example 1: Simulation

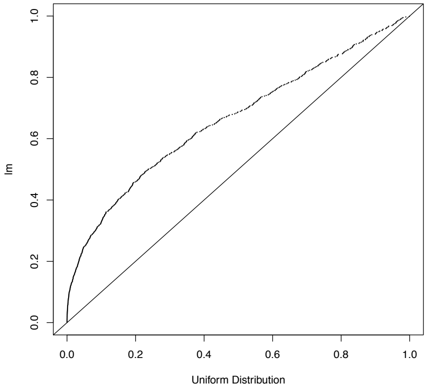
- Simulation outline
    1. Simulate a data set
    2. Test $H_0 : \mu = 0$ ignoring the random effect (save $P$-value)
    3. Test $H_0 : \mu = 0$ accounting for the random effect (save $P$-value)
- Repeat the three steps 999 additional times
- Given that the *true* $\mu = 0$ (by design), we would expect 50 of these $P$-values to be less than 0.05
- Why?

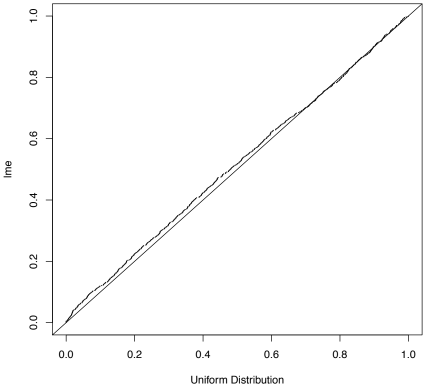
# Example 1: Results

```r
set.seed(210)
res = replicate(B3, sim.ranef(3, 6, 0.25, 0.5, verbose = FALSE))
mean(res[1, ] < 0.05)


## [1] 0.247


mean(res[2, ] < 0.05)


## [1] 0.072
```

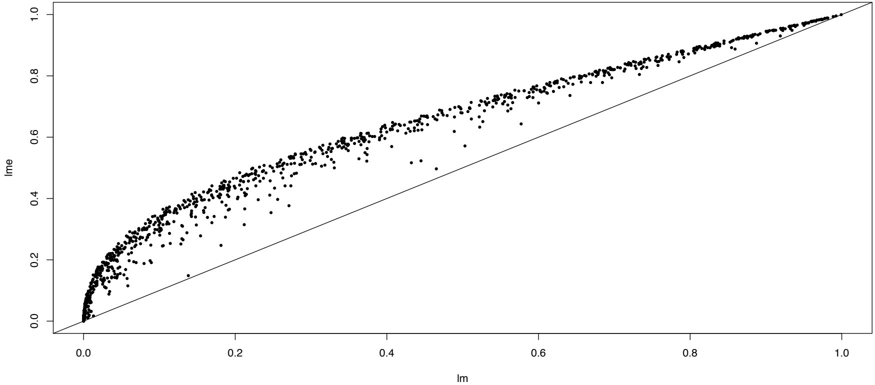
- The empirical type I error rate when not accounting for the random effect is 0.25.
- This inflated by a factor of 4.9.
- The empirical error rate when accounting for the random effect is slightly inflated
- This is due to the small sample size ($n = 6$)
- More on this later.

# Example 1: Results





# Example 1: Results



# Example 1: Results

- Now, we repeat the simulation with a larger sample size

```
res = replicate(B3, sim.ranef(3, 50, 0.25, 0.5, verbose = FALSE))
mean(res[1, ] < 0.05)


## [1] 0.215


mean(res[2, ] < 0.05)


## [1] 0.052
```

- The empirical type I error when not accounting for the random effect remains inflated by a factor of 4.3.
- The empirical type I error when accounting for the random effect is now right about the nominal level of 0.05

# Example 2: Setup

- Now consider the two-sample problem we have previously considered with a twist
- Question: Does treatment alter the distribution of the RNA level of a given gene?
- Assumptions:
    - the RNA level for the untreated group follows a normal distribution with mean $\mu_0$ and variance $\sigma^2$
    - The RNA level for the treated group follows a normal distribution with mean $\mu_1$ and variance $\sigma^2$
- Sample $n$ units from each treatments in replicates of 3
- Apply the two-sample t-test which does not account for the clustering

# Example 2: Simulation

```r
set.seed(2314)
# Simulate with no clustering effect (sb=0)
pval0 = replicate(B3, sim.twosample.clustered(3, 10, 0.25, 0))
# Simulate with no clustering effect (sb>0)
pval1 = replicate(B3, sim.twosample.clustered(3, 10, 0.25, 0.5))
mean(pval0 < 0.05)
```

```
## [1] 0.049
```

```r
mean(pval1 < 0.05)
```

```
## [1] 0.252
```

- The empirical type I error when there is no clustering effect is 0.049
- The empirical type I error when there is a clustering effect is 0.25
- This off by a factor of 5!