Statistics

Biostatistics and Bioinformatics

June 17-18, 2021

Modeling Two Populations

- Suppose we are interested in the **population mean** of the expression level of a gene in two populations: parental cell line vs UV2
- We **randomly** select experimental units from each population and conduct RNA-seq experiment
- Denote the population mean expression by μ_1 for parental and μ_2 for UV2
- Modeling the observed gene expression Y
  - If unit i is parental: $Y_i = \mu_1 + \epsilon_i$
  - If unit i is UV2: $Y_i = \mu_2 + \epsilon_i$
  - A unified formula: denote $Z_i = 0$ (or $Z_i = 1$) if unit i is parental (or UV2)

$$Y_i = \mu_1 + (\mu_2 - \mu_1)Z_i + \epsilon_i = \beta_0 + \beta_1 Z_i + \epsilon_i$$

Assumptions

$$Y_i = \mu_1 + (\mu_2 - \mu_1)Z_i + \epsilon_i = \beta_0 + \beta_1 Z_i + \epsilon_i$$

- What we observe is a perturbed value of the population means (what happens if $\epsilon_i = 0$?)
- Independence: The noise ϵ are independent across units i
- Normality: The noise ϵ follows $N(0, \sigma^2)$, i.e. normal distribution with mean 0 and variance σ^2 for all units i
- Homoskedasticity: the noise level σ^2 is the same for all units i

Interpretation

The unified formula: denote $Z_i = 0$ (or $Z_i = 1$) if unit i is parental (or UV2)

$$Y_i = \mu_1 + (\mu_2 - \mu_1)Z_i + \epsilon_i = \beta_0 + \beta_1 Z_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

- The binary valued Z_i is also known as dummy variable
- $\beta_0 = \mu_1$ is the population mean of the baseline group (parental in this case)
- $\beta_0 + \beta_1$ is the population mean of the other group (UV2 in this case)
- β_1 is the difference in the population means between the two groups (what does it mean when $\beta_1 > 0, = 0 <, 0$?)

Signal vs Noise

$$Y_i = \mu_1 + (\mu_2 - \mu_1)Z_i + \epsilon_i = \beta_0 + \beta_1 Z_i + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2)$$

- When the noise level σ^2 is larger, the observed expression contains more uncertainty (i.e. noisier), making it harder to estimate the population means
- The magnitude of noise level σ^2 is relative to signals β_0 and β_1 (or equivalently μ_1 and μ_2)

$$1000\,Y_i = 1000\beta_0 + 1000\beta_1 Z_i + 1000\epsilon_i$$

- The value of σ^2 is unknown, but is important for the inference of population means

Modeling Two Binary Factors

- Suppose we have two factors: cell line (parental vs UV2) and treatment (isotype vs anti-PD1)
- We are interested in the effect of those two factors on the **population mean** of the expression level of a gene
- How many populations do we have?
- We **randomly** select experimental units from each population and conduct RNA-seq experiment
- Denote the population mean expression by
  - μ_{11} for parental+isotype
  - μ_{21} for parental+anti-PD1
  - μ_{12} for UV2+isotype
  - μ_{22} for UV2+anti-PD1
- For each experimental unit, we observe $Y_i = \mu_{??} + \epsilon_i$

A Unified Formula for Two Factors

- Denote $Z_i = 0$ (or $Z_i = 1$) if unit i is parental (or UV2) cell line
- Denote $X_i = 0$ (or $X_i = 1$) if unit i is treated by isotype (or anti-PD1)
- parental is the baseline for cell line, and isotype is the baseline for treatment

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i + \epsilon_i$$

	parental	UV2
isotype	$\mu_{11} = \beta_0$	$\mu_{12} = \beta_0 + \beta_2$
anti-PD1	$\mu_{21} = \beta_0 + \beta_1$	$\mu_{22} = \beta_0 + \beta_1 + \beta_2 + \beta_3$

Interpretation

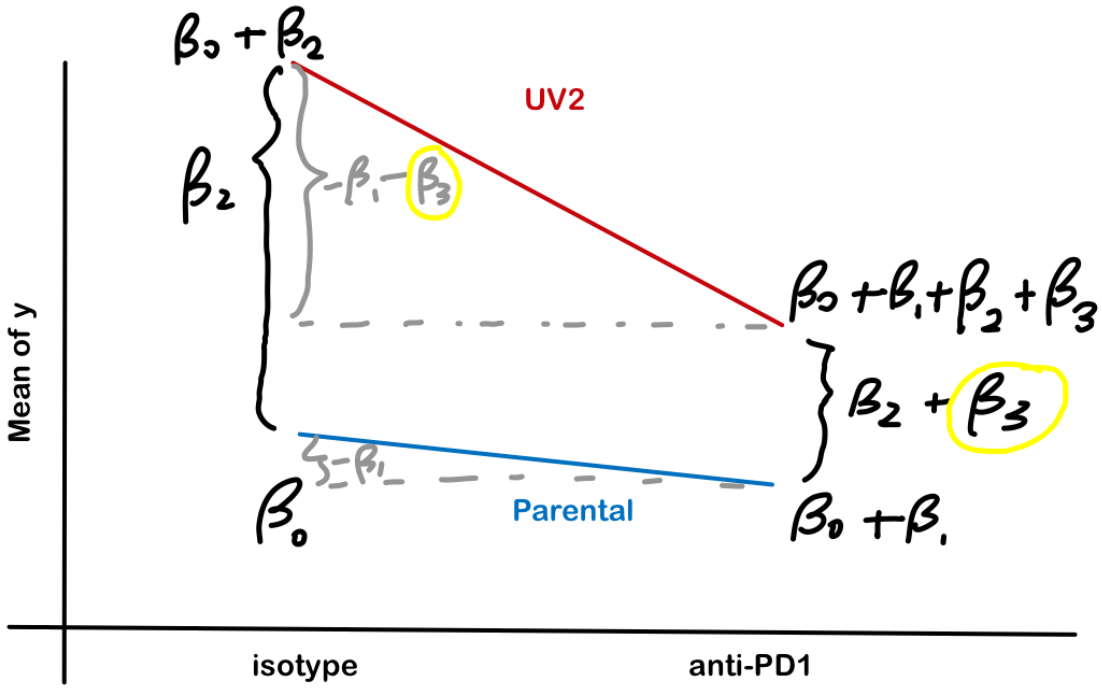
	parental	UV2	difference
isotype	β_0	$\beta_0 + \beta_2$	β_2
anti-PD1	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_2 + \beta_3$
difference	β_1	$\beta_1 + \beta_3$	β_3

- β_0: mean expression level of the baseline population
- β_1: the difference between the two treatments in parental (baseline) cell line
- β_2: the difference between the two cell lines when treated by isotype (baseline)
- $\beta_1 + \beta_3$: the difference between the treatments in UV2 cell line
- $\beta_2 + \beta_3$: the difference between the cell lines when treated by anti-PD1
- β_3: interaction effect between cell line and treatment
- What happens when $\beta_3 = 0$?

Interaction Effect

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i + \epsilon_i$$

- $X_i = 0$ if treatment is isotype, $X_i = 1$ if treatment is anti-PD1.
- $Z_i = 0$ if cell line is parental, $Z_i = 1$ if cell line is UV2.

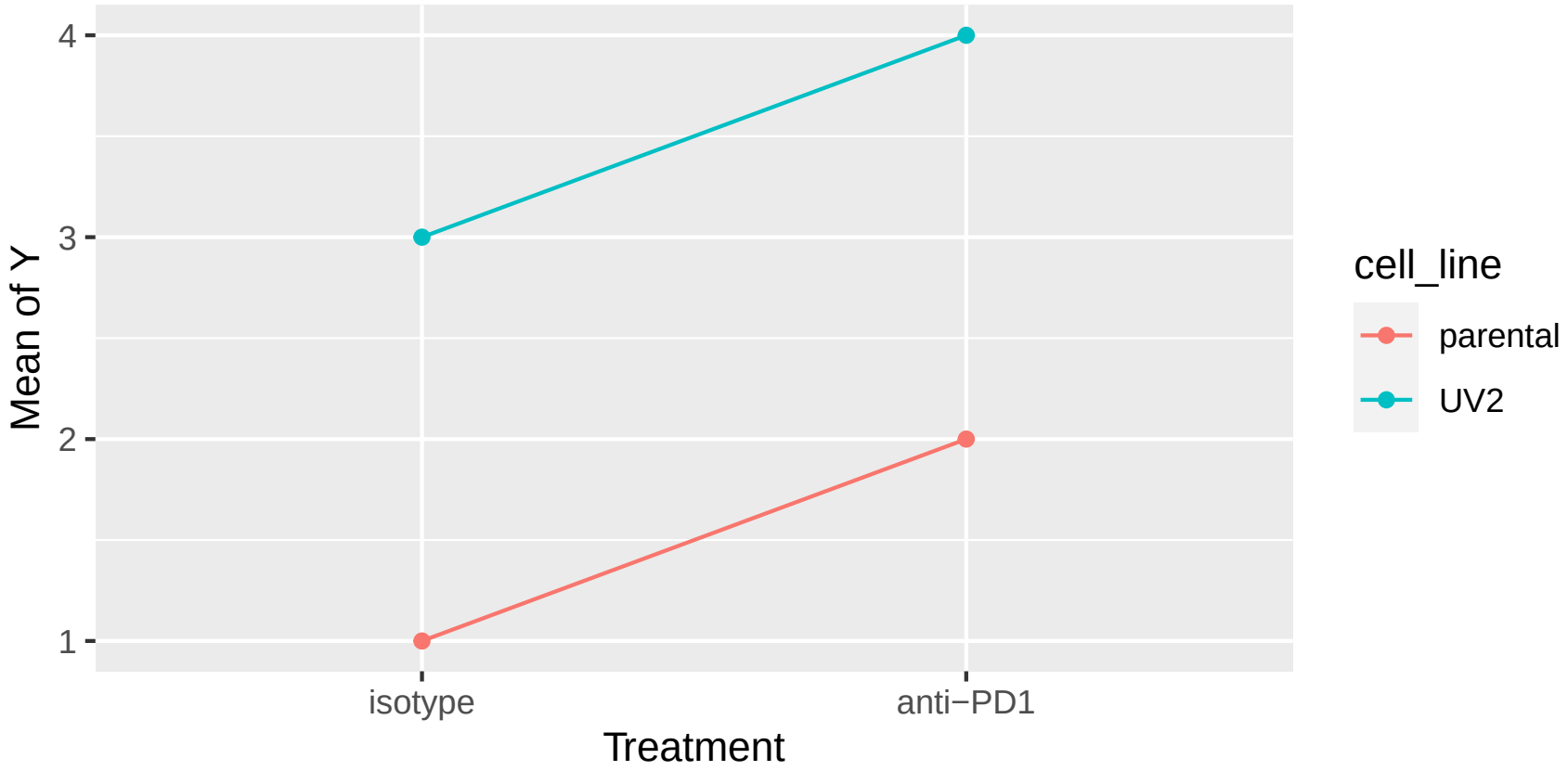


- Pay attention to the signs of βs!

Example - No Interaction

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i + \epsilon_i$$

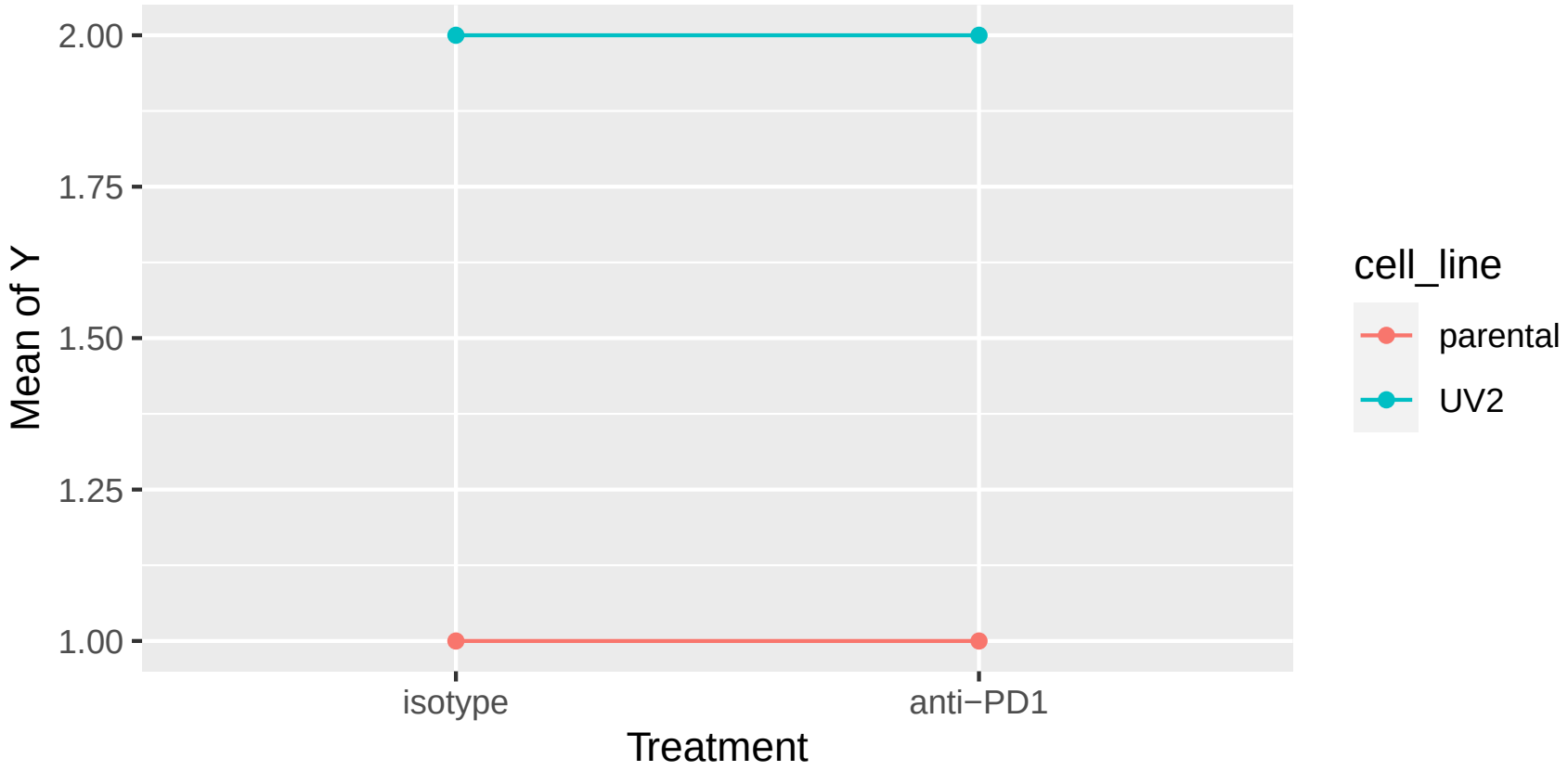
X_i for treatment, Z_i for cell line



Example - No Interaction

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i + \epsilon_i$$

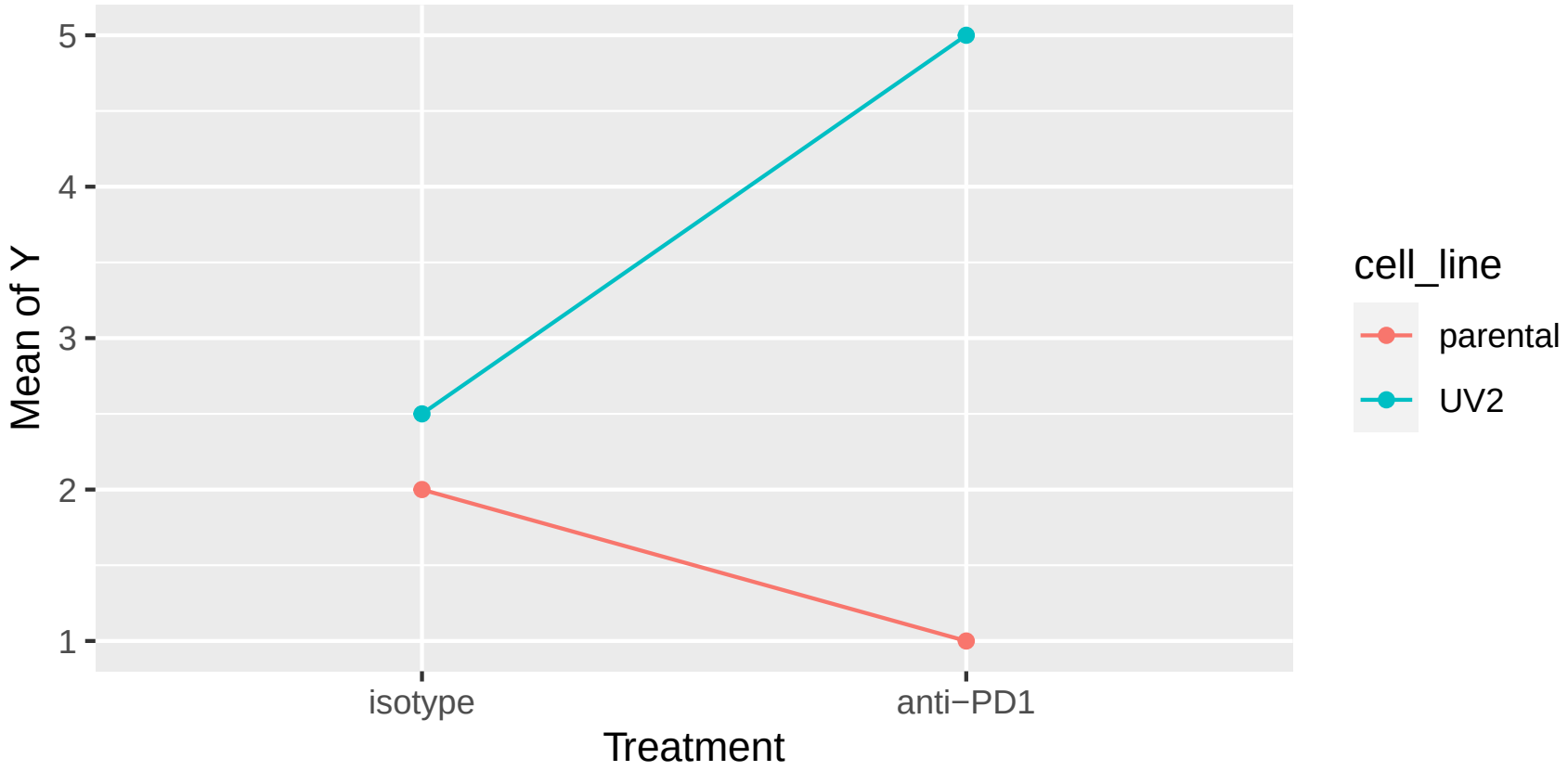
X_i for treatment, Z_i for cell line



Example - With Interaction

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i + \epsilon_i$$

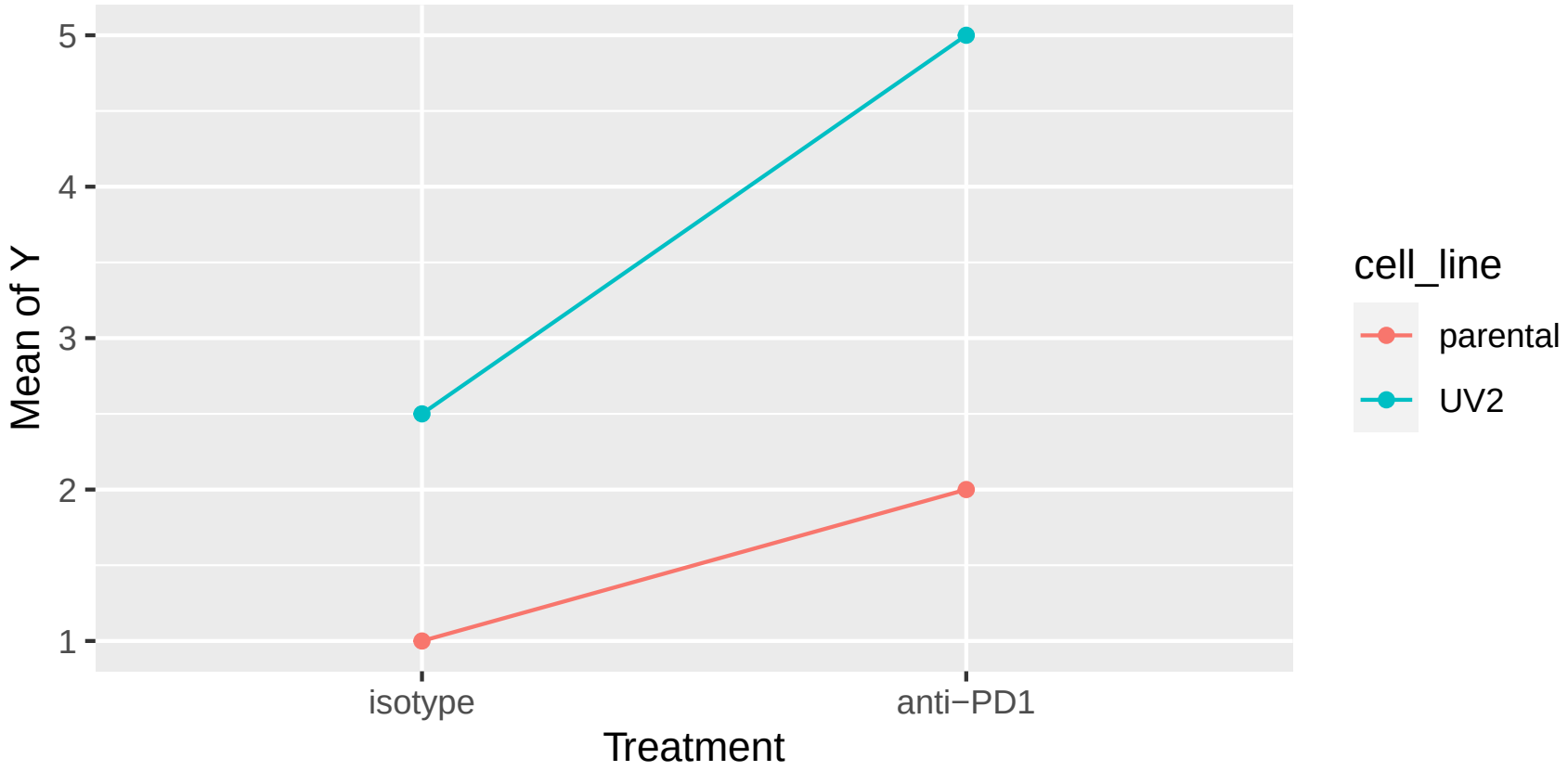
X_i for treatment, Z_i for cell line



Example - With Interaction

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i + \epsilon_i$$

X_i for treatment, Z_i for cell line



Interaction $\neq$ Correlation

Using treatment (isotype, anti-PD1) and cell line (parental, UV2) as example:

- Correlation:
  - The treatment assignment is dependent on the cell line.
  - It has nothing to do with Y
- Interaction
  - The effect of treatment on Y is different between cell lines
  - Assignment of treatment may or may not depend on cell line

Point and Interval Estimate

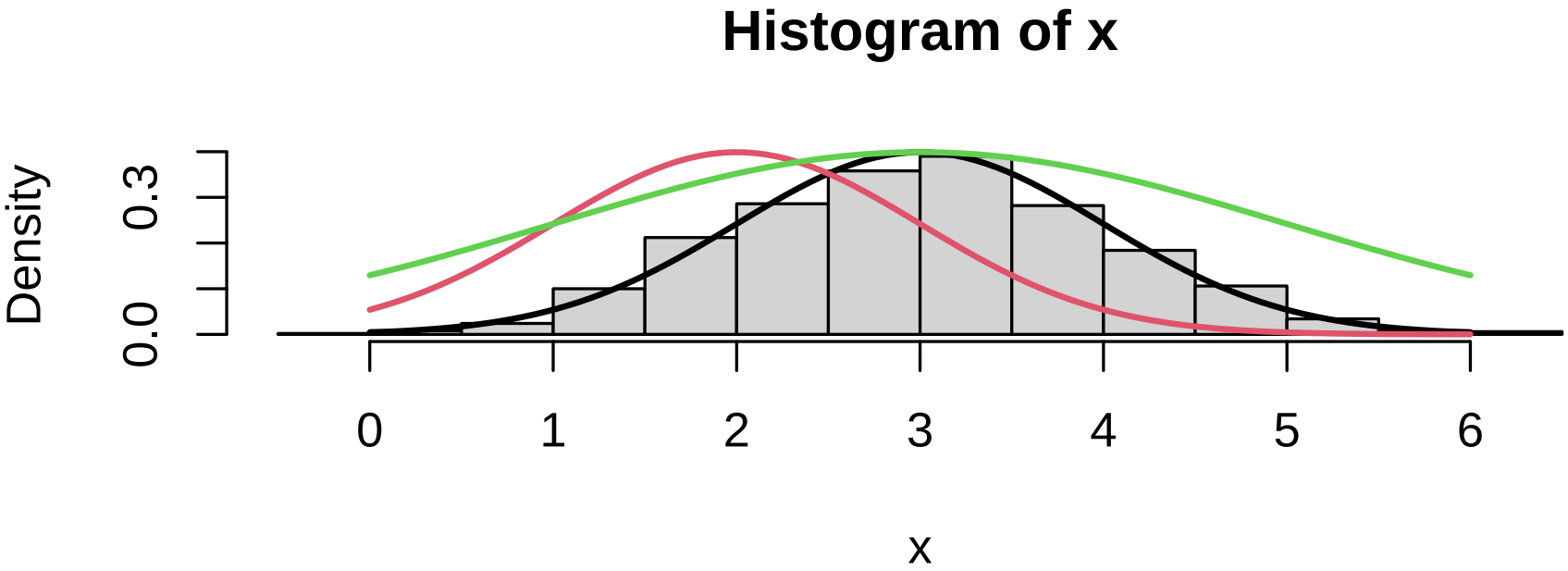
- We are often interested in estimating the unknown parameters in a model
- Example using the two-population model:

$$Y_i = \mu_1 + (\mu_2 - \mu_1)Z_i + \epsilon_i = \beta_0 + \beta_1 Z_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

  - Mean level for parental cell line: μ_1
  - Mean level for UV2 cell line: μ_2
  - Fold-change: $\rho = \mu_1/\mu_0$
  - Standardized difference: $\Delta = |\mu_1 - \mu_2|/\sigma = |\beta_1|/\sigma$
- Two types of estimation
  - Point estimation
  - Interval estimation

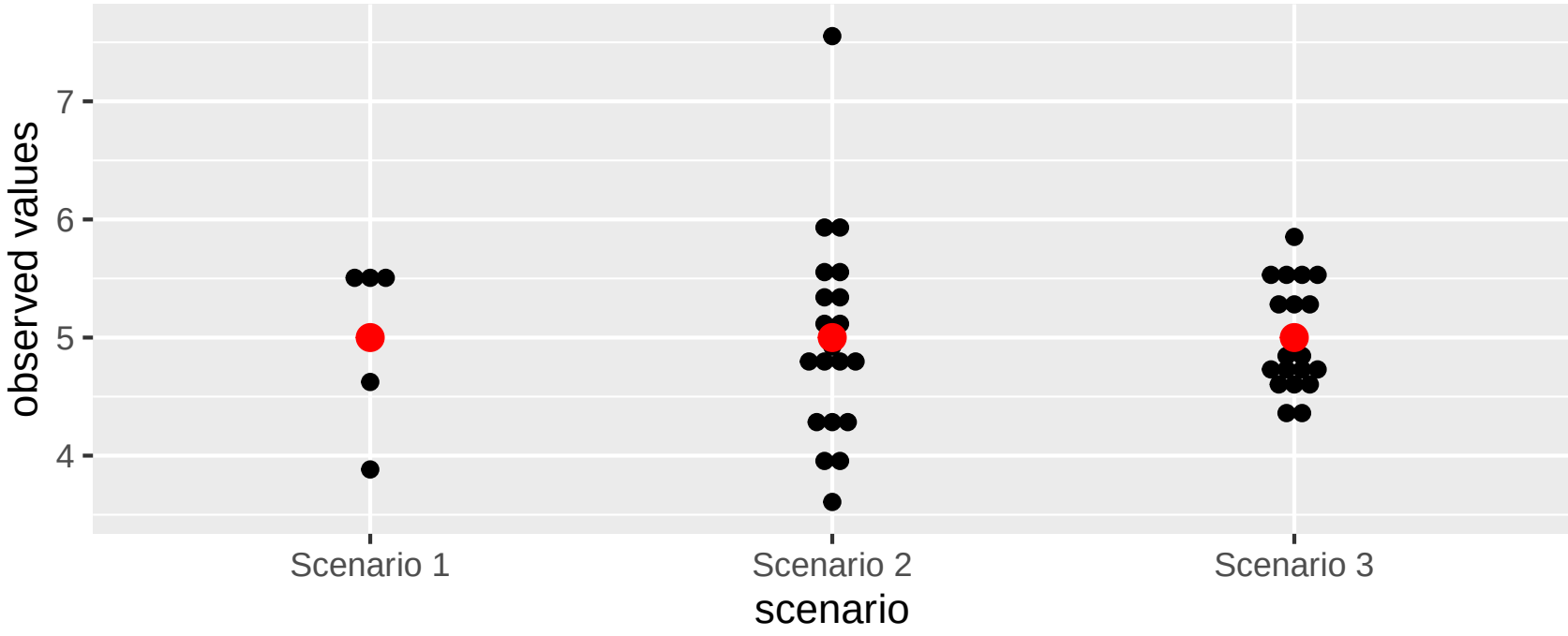
Point Estimation

- We can estimate the parameters with anything
  - Sample mean of each cell line for μ_1 and μ_2
  - Sample median, less prone to the influence of outliers
  - You can estimate μ_1 using $\widehat{\mu}_1 = 0$
- A popular method is the maximum likelihood estimator (MLE)
  - Idea: look for the value of parameter such that observing the observed values is most likely



Limitation of Point Estimate

- It is either equal to the true value of parameter or is not
- With a single number, we don't know how accurate it is
- The accuracy of the point estimator is generally affected by
  - Sample size: larger sample size $\Rightarrow$ more accurate
  - Noise level: larger noise level $\sigma \Rightarrow$ less accurate



Confidence Interval (CI)

- The accuracy of each point estimate is quantified by its margin of error
- The confidence interval is usually obtained as the point estimate plus or minus the margin of error
- The width of the confidence interval is affected by
  - The coverage probability: Larger coverage $\Rightarrow$ longer CI
  - Sample size: larger sample size $\Rightarrow$ shorter CI
  - Noise level: larger noise level $\sigma \Rightarrow$ longer CI
- $(-\infty, +\infty)$ is also a CI with coverage probability 100%, but is not meaningful
- CI can be one-sided or two-sided
  - Drug efficacy example: you want to put a lower limit to the improvement caused by the drug
  - When the coverage probability is the same, one-sided limits are generally tighter than two-sided limits

Interpretation of Confidence Interval

Using the two-population example:

- Suppose we're interested in the differential expression between cell lines $\beta_1 = \mu_2 - \mu_1$
- Based on the data, a stat software produces a 95% CI $[0.3, 0.8]$ for β_1
- Wrong statement: β_1 is covered between 0.3 and 0.8 with probability 0.95
  - β_1, 0.3, 0.8 are all fixed numbers, β_1 is either covered or not covered
- Correct statement: we are 95% "confident" that β_1 is covered by our CI
  - If we redo the entire experiment by sampling from the same population for N rounds
  - For each round, we calculate the CI using the same fashion. β_1 is either covered or not covered by the CI in that round
  - If we count the number of rounds with β_1 covered by the CI, it will roughly be $N * 0.95$

A Simulated Example for Two Populations

```r
# mu1=3, mu2=4, beta1=1
y <- c(rnorm(10)+3, rnorm(10)+4)
x <- c(rep(0,10),rep(1,10))
lm_res <- lm(y~x)
print(confint(lm_res))
```

```
##                  2.5 %   97.5 %
## (Intercept)  2.373671 3.748822
## x           -0.465177 1.479581
```

```r
L <- confint(lm_res)[2,1]
U <- confint(lm_res)[2,2]
print(L < 1 & 1 < U)
```

```
## [1] TRUE
```

A Simulated Example for Two Populations

```r
covered <- rep(FALSE, 1000)
for (i in 1:1000){
  y <- c(rnorm(10)+3, rnorm(10)+4)
  x <- c(rep(0,10),rep(1,10))
  lm_res <- lm(y~x)
  L <- confint(lm_res)[2,1]
  U <- confint(lm_res)[2,2]
  covered[i] <- (L < 1 & 1 < U)
}
table(covered)
```

```
## covered
## FALSE  TRUE
##    49   951
```

Choice of Null and Alternative Hypotheses

- H_0 denotes null hypothesis, H_a or H_1 denotes alternative hypothesis.
- H_0 posits the status quo
- H_0 is the conservative hypothesis
- In the US legal system, the defendant is assumed to be innocent. H_0: innocence
- Drug safety study:
  - H_0: Drug is toxic
  - H_1: Drug is safe
- Drug efficacy study:
  - H_0: Drug is not efficacious
  - H_1: Drug is efficacious
- Setup of hypotheses require expressing your scientific question using parameters in the model

Hypotheses in the Two-Population Example

- $Z_i = 0$ (or $Z_i = 1$) if unit i is parental (or UV2), we observe

$$Y_i = \mu_1 + (\mu_2 - \mu_1)Z_i + \epsilon_i = \beta_0 + \beta_1 Z_i + \epsilon_i$$

- H_0: The population mean of two cell lines are the same
  - $\mu_1 = \mu_2$, or equivalently $\beta_1 = 0$
- H_1 can have several setups:
  - The population mean of two cell lines are different: $\mu_1 \neq \mu_2$, or $\beta_1 \neq 0$
  - UV2 expresses more than parental cell line: $\mu_1 > \mu_2$, or $\beta_1 > 0$
  - UV2 expresses less than parental cell line: $\mu_1 < \mu_2$, or $\beta_1 < 0$

Hypotheses in the Two-Factor Example

When no interaction is modeled:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \epsilon_i$$

- Testing the treatment effect
  - $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$, no treatment effect
  - $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 > 0$, baseline treatment express less
  - $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 < 0$, baseline treatment express more
- Testing the cell line effect
  - $H_0 : \beta_2 = 0$ vs $H_1 : \beta_2 \neq 0$, no cell line effect
  - $H_0 : \beta_2 = 0$ vs $H_1 : \beta_2 > 0$, baseline cell line express less
  - $H_0 : \beta_2 = 0$ vs $H_1 : \beta_2 < 0$, baseline cell line express more

Hypotheses in the Two-Factor Problem

With interaction is modeled:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i + \epsilon_i$$

- H_0: there is no interaction effect between cell line and treatment on the expression level, $\beta_3 = 0$
- H_1: there is an interaction effect, $\beta_3 \neq 0$
- Testing for cell line (or treatment) effect is not the same as testing β_1 (or β_2)!
  - $H_0 : \beta_1 = 0$ test for the treatment effect in parental cell line
  - $H_0 : \beta_2 = 0$ test for the cell line effect under isotype treatment
  - H_1 can be one-sided or two-sided
  - We will not discuss the hypotheses for test of treatment or cell line effect here

Testing Hypotheses

- A test is a decision rule based on the data: if data follow a given pattern, reject H_0, otherwise, do not reject H_0
- Example of a decision rule: if the difference in sample means between the two cell lines is larger than 2, reject $H_0 : \mu_1 = \mu_2$
- Example of other decision rules:
  - Always reject H_0
  - Always don't reject H_0
  - Flip a coin, if head, reject H_0; if tail, don't reject H_0

Type I error and Type II error

For any decision rule, due to the randomness in the data, we may make two types of errors.

	H_0 is true	H_0 is false
H_0 is not rejected	correct inference true negative (TN)	**type II error** false negative (FN)
H_0 is rejected	**type I error** false positive (FP)	correct inference true positive (TP)

- α: probability of type I error
- β: probability of type II error

Type I and Type II Error Trade-off

	H_0 is true	H_0 is false
H_0 is not rejected	correct inference	**type II error**, β
H_0 is rejected	**type I error**, α	correct inference

- There is a trade-off between α and β
- In statistics, we control α below a certain level (e.g. 0.05), and use decision rule with small β. This is related to setting H_0 as the conservative hypothesis
- Drug efficacy example: H_0 : not efficacious, H_1 : efficacious
  - Type I error: conclude that the drug is efficacious when it's not
  - Type II error: conclude that the drug is not efficacious when it is

Meaning of Rejection and No Rejection

- Since type I error and type II error are not treated symmetrically, rejection and no rejection of H_0 are also not symmetric
- Rejection of H_0 means: I have so much evidence to reject H_0 that the probability of a wrong rejection is $< \alpha$
- No rejection of H_0 means: I don't have enough evidence to reject H_0 while keeping the probability of wrong rejection $< \alpha$. It doesn't mean we have enough evidence that H_0 is true. If I say H_0 is true, in most cases I won't be able to tell you the probability of making a wrong acceptance of H_0 (type II error).

Factors Affecting Power

- α: probability of type I error
- β: probability of type II error
- $1 - \beta$: **power** of the decision rule (power of the test)
- The power of tests such as t tests and Z tests are often affected by several factors. Using the two-population example:
  - $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$
  - Effect size: larger $|\mu_1 - \mu_2| \Rightarrow$ larger power
  - Noise level: large noise level $\sigma^2 \Rightarrow$ smaller power
  - Sample size: larger sample size $\Rightarrow$ larger power
- We won't know the value of β in this case because $\mu_1 - \mu_2$ is unknown

P-Value

- p-value is the probability of observing data "more extreme" that what you currently have if H_0 is true
- "more extreme" is defined corresponding to the direction of H_1:
  - Denote $\bar{y}_1$ and $\bar{y}_2$ as the sample means of the two cell lines
  - $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$, more extreme means $|\bar{y}_1 - \bar{y}_2|$ is larger
  - $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 < \mu_2$, more extreme means $\bar{y}_1 - \bar{y}_2$ is smaller
  - $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 > \mu_2$, more extreme means $\bar{y}_1 - \bar{y}_2$ is larger
- p-value is smaller when the observed data is more in the direction of H_1 compared to what is expected when H_0 is true. So reject H_0 and accept H_1 when p-value is very small.

P-Value

- We often reject H_0 when p-value< 0.05
  - This is a decision rule with probability of type I error < 0.05
  - There is nothing special about $\alpha = 0.05$. The choice of α reflects your tolerance on type I error
- Smaller p-value means that the effect is so significant that you can reject H_0 with smaller probability of making type I error

P-Value and Effect Size

- Effect size is relative to the noise level

$$Y_i = \beta_0 + \beta_1 Z_i + \epsilon_i, \quad 1000\, Y_i = 1000\beta_0 + 1000\beta_1 Z_i + 1000\epsilon_i$$

- Larger effect size $\not\Rightarrow$ smaller p-value. Two genes with the same β_1 may have different variance of ϵ_i, i.e. σ^2. The gene with smaller σ^2 will have smaller p-value.
- Both effect size and p-value are important:
  - A drug improving survival by 1 day with a p-value of 0.01
  - A drug improving survival by 10 years with a p-value of 0.05

Connection Between CI and Hypotheses Testing

Using $\beta_1 = \mu_2 - \mu_1$ in the two-population example to illustrate:

- A CI for β_1 with coverage probability $1 - \alpha$ corresponds to a test for β_1 with type I error probability $< \alpha$
- Two-sided CI $[L, U]$ can be used to test for $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$
- One-sided CI $[L, \infty]$ can be used to test for $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 > 0$
- One-sided CI $[-\infty, U]$ can be used to test for $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 < 0$