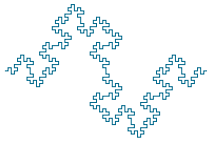
High-Throughput Sequencing Course

Statistical Inference: Part I

Biostatistics and Bioinformatics



Summer 2019

Duke University School of Medicine

Anatomy of a Simple Model: Population

- We are interested in the expression level of a gene in a population
- Denote that level by μ
- Some questions
  - Is $\mu > 0$ (over expressed)
  - What is the value of μ
- Model the observed gene expression Y as

$$Y = \mu + \epsilon$$

- ϵ is a random error term
- In English:
  - In absence of random noise ($\epsilon = 0$), we observe the true value μ
  - In practice: We observe a perturbed value of μ (the truth plus random error)
- What are the assumptions thus far?

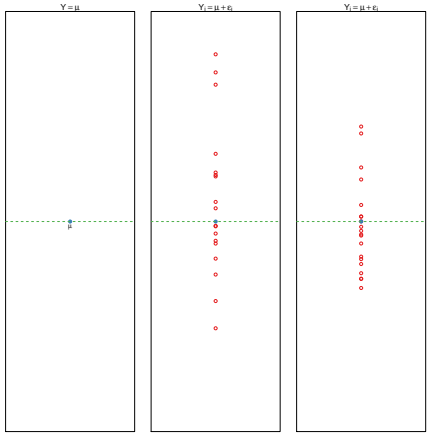
Anatomy of a Simple Model: Sampling from Population

- Our design: Sample n experimental units from this population
- What are observed are $Y_1, \ldots, Y_n$
- Model the observed gene expression for the i-th experimental units as

$$Y_i = \mu + \epsilon_i$$

- Note that we are not interested in any of n experimental units
- We are interested in the population
- Note that Y_i and ϵ_i are indexed by i
- μ (population parameter) is *not* indexed by i.

Illustration of Model



Anatomy of a Simple Model: Assumptions

- The observed expression (Y_i) is signal (μ) plus noise (ϵ_i)
- Additional assumptions:
  - The error terms for the experimental units $\epsilon_1, \ldots, \epsilon_n$ are mutually independent
  - The error terms $\epsilon_1, \ldots, \epsilon_n$ are identically distributed
  - More specifically: The error terms follow a normal distribution with mean 0 and variance σ^2.
- Preview: Models for analysis of RNA-Seq data will also involve nuisance parameters
- While there are of no direct interest, they have to be dealt with appropriately in any rigorous and principled analysis

Anatomy of a Simple Model: A Summary

- Y_i is random but observable outcome
- ϵ_i is a random but unobservable outcome
- μ is the unknown population parameter of interest
- $\epsilon_1, \ldots, \epsilon_n$ are mutually independent
- $\epsilon_1, \ldots, \epsilon_n$ are normally distributed with mean 0 and *common* variance σ^2
- σ^2 is an unknown population nusiance parameter of interest

Anatomy of a Simple Model: Final Notes

- In statistics, we base the inference on the observations Y_i
- ϵ_i is called a latent variable
- The estimation of σ^2 is a nuisance that has to be dealt with
- It is not realistic to assume that we know what σ^2 is.
- The variability of estimating σ^2 has to be accounted for in the inference
- There is no point to statistics if we assume we know μ. Why?

Statistical Hypothesis Testing: A Generic Overview

- Formulate a scientific hypothesis
- Formulate the corresponding statistical hypothesis
- This will consist of a *null* hypothesis (H_0) and an *alternative* hypothesis (H_1)
- Specify an experimental design
- Specify the testing procedure to be used:
  - an appropriate test statistic
  - decision rule based on the test statistic (typically under a set of assumptions)
- Execute Experiment (collect data)
- Based on the amount of evidence using the decision rule
  - either conclude there is evidence to reject the null hypothesis H_0 in favor of H_1
  - or fail to reject H_0 (inconclusive)

IMPORTANT: Failing to reject H_0 does *not* afford us to conclude that H_0 is *true*

Experimental Design

- Two examples
  - Decide upfront to evaluate $n = 10$ experimental units
  - Decide to initially evaluate $n_1 = 5$ experimental units (Stage 1). Depending on the results evaluate an additional $n_2 = 5$ experimental units
- These are *different* experimental strategies
- Design 1: The final sample size is $n = 10$
- Design 2: The final sample size is $n = n_1 = 5$ *or* $n = n_1 + n_2 = 10$
- The statistical properties of your decision rule depends on the strategy used.
- More on design of experiments (DOE) in Week 3

Null versus Alternative

- The null hypothesis posits the status quo
- It is the conservative hypothesis
- In the US legal system, the defendant is assumed to be innocent
- The null hypothesis: Defendant is innocent
- Study: Investigate if gene XYZ is differentially expressed with respect to treatment
- In other words, does the distributions of the feature of the gene you are interested in change when the experimental unit is exposed to treatment?
  - H_0 gene XYZ is *not* differentially expressed with respect to treatment
  - H_1 gene XYZ is differentially expressed with respect to treatment

More on Null versus Alternative

- Suppose that your are studying the effect of a drug in a clinical study
- Safety Study:
  - H_0: Drug is toxic
  - H_1: Drug is safe
- Efficacy study:
  - H_0: Drug is not efficacious
  - H_1: Drug is efficacious

Notation: True versus False Null Hypothesis

- The truth may be stated either by the null or alternative hypothesis
- If the truth is stated by the statement of the null hypothesis, we will say that
  - The null hypothesis is true
  - or call it a true null hypothesis
- If the truth is stated by the statement of the alternative hypothesis, we will say that
  - The null hypothesis is false
  - or call it a false null hypothesis
- We will use these terms for notational convenience

Type I and II errors

- Type I Error: Erroneously decide in favor of the alternative hypothesis (reject a true null hypothesis)
- Type II Error: Erroneously decide in favor of the null hypothesis (fail to reject a false null hypothesis)
- The so called "alpha" level is the probability of a type I error
- The "power" of a test, is the complement of the probability of the type II error
- IMPORTANT: There is a trade-off between these two error rates

Type I and II errors

		Null Hypothesis (H_0)	
		True	**False**
Test Decision	**Reject (*Significant p*)**	Type I error (α)	Correct inference
	Fail to reject (*Not significant p*)	Correct inference	Type II error (β)

$$\text{Power} = 1 - \beta$$

Type I and II error trade-off

- In our court system, a defendant is assumed innocent until proven guilty
  - Type I error: Convict an innocent defendant
  - Type II error: Free a guilty defendant
- If the prosecution gets too much leeway, the the likelihood of convicting an innocent defendant increases
- Conversely, if the prosecution is reigned in by the judge, the likelihood of letting a guilty defendant walk free increases
- Similar analogy in the case of a smoke detector
  - Dialing up the sensitivity, increases the likelihood of annoying beeps when using your toaster
  - Dialing down the sensitivity, increases the likelihood of missing a true fire

Notation: Decision

- false-positive (FP): Reject a true null hypothesis (Type I error)
- true-positive (TP): Reject a false null hypothesis
- false-negative (FN): Fail to reject a false null hypothesis (Type II error)
- true-negative (TN): Fail to reject a false null hypothesis
- We will use these terms for notational convenience

		Null Hypothesis (H_0)	
		True	**False**
Test Decision	**Reject (*Significant p*)**	Type I error (α) False Positive (FP)	Correct inference True Positive (TP)
	Fail to reject (*Not significant p*)	Correct inference True Negative (TN)	Type II error (β) False Negative (FN)

Three Decision Rules

- Following the collection of data, consider using one of the three decision rules
- Decision Rule 1: Reject H_0
- Decision Rule 2: Do not reject H_0
- Decision Rule 3: Flip a coin: Reject H_0 if tails and do not reject H_0 if heads
- What are the type I and II error rates for these decision rules?
- Which one would you choose?

Decision Rule 1

- Decision: Reject H_0
- If H_0 is true, then it will be rejected
- A false-positive decision will be made if H_0 is true
- $\alpha = 1$
- If H_0 is false, then it will be rejected
- A true-positive decision will be made if H_0 is false
- $\beta = 0$

Decision Rule 2

- Decision: Do not reject H_0
- If H_0 is true, then it will not be rejected
- A false-positive decision will not be made
- $\alpha = 0$
- If H_0 is false, then it will not be rejected
- A false-negative decision is will be made
- $\beta = 1$

Decision Rule 3

- Decision: Flip a coin: Reject H_0 if tails and do not reject H_0 if heads
- If H_0 is true, then the probability of rejecting it is one-half
- $\alpha = \frac{1}{2}$
- If H_0 is false, then probability of not rejecting it is one-half
- $\beta = \frac{1}{2}$

A Bad Rule is a Valid (but bad) Decision Rule

- Decision Rule 1: Reject H_0
  - $\alpha = 1$ and $\beta = 0$
- Decision Rule 2: Do not reject H_0
  - $\alpha = 0$ and $\beta = 1$
- Decision Rule 3: Flip a coin: Reject H_0 if tails and do not reject H_0 if heads
  - $\alpha = \frac{1}{2}$ and $\beta = \frac{1}{2}$
- Note that these decision rules effectively ignore the data
- While they are poor decision rules, they are technically valid decision rules
- A poor statistical approach will effectively reduce to one of the three
- Note that while $\alpha + \beta = 1$ in all these cases, that is generally not the case
- The type I error is generally *not* the complement of the type II error

A Simple Example: Formulation

- You suspect that a coin (H on side and T on the other) is not fair (biased)
- Let π denote the probability that the coin lands a head on any given toss
- A coin is "fair" if $\pi = \frac{1}{2}$
- or is "biased" otherwise (i.e., $\pi \neq \frac{1}{2}$)
- It is more likely to land a tail if $\pi < \frac{1}{2}$
- or more likely to land a head if $\pi > \frac{1}{2}$

A Simple Example: Statistics and Plain English

- The statistical hypotheses could be succinctly stated as:
  - Test $H_0 : \pi = \frac{1}{2}$ against $H_1 : \pi \neq \frac{1}{2}$
- In English:
  - We give benefit of the doubt to the fact that the coin is fair and then will, under this assumption, ascertain if there is enough evidence, on the basis of the data, to conclude that the coin is biased

A Simple Example: Decision Rule

- Following the formulation of the hypotheses, we have to decide on an experimental design and a decision rule
- These, along with the specification of the hypotheses, should be done before collecting data. Why?
- Our experimental design: flip the coin $n = 12$ times
- Why $n = 12$ and not say $n = 13$ (more on this later)
- A reasonable decision rule for this type of design is to use the so called Binomial Test
- We will skip the technical details on the test

A Simple Example: Collect Data

- We conduct the experiment and observe

```
##  [1] "T" "T" "T" "T" "T" "H" "T" "H" "T" "T" "T" "T"
```

- There are (per design) 12 flips of the coin
- We observe 2 heads and 10 tails
- What would you conclude?
- Would you reject if the number of heads were 3?
- how about 4?
- or 5?

A Simple Example: Binomial Test in Action

- We conduct the binomial test

```
##
##  Exact binomial test
##
## data:  sum(x == "T") and length(x)
## number of successes = 10, number of trials = 12, p-value = 0.03857
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.5158623 0.9791375
## sample estimates:
## probability of success
##              0.8333333
```

- What should we conclude?
- At the $\alpha = 0.05$ level, there is sufficient evidence to reject the hypothesis that the coin is fair (P-value=0.039)
- Note that there is *not* sufficient evidence to reject the null if you wish to control the type I error rate at $\alpha = 0.01$

Anatomy of a Two-sample Model

- Scientific Hypothesis: Treatment affects level of the gene
- Let μ_0 denote the population level for the gene of interest in the *untreated* population
- Let μ_1 denote the population level for the gene of interest in the *treated* population
- $\mu_0 = \mu_1$: There is no treatment effect on the gene
- $\mu_0 \neq \mu_1$: There is a treatment effect (gene level is "differentiably" expressed with respect to treatment)
- $\mu_0 < \mu_1$: treatment increases level of gene
- $\mu_0 > \mu_1$: treatment decreases level of gene

Two-sample Model: Observations

- Observation from Untreated Population:

$$X = \mu_0 + \epsilon$$

- Observation from Treated Population:

$$Y = \mu_1 + \epsilon'$$

- For each population, the observed mRNA level is the population level (μ_0 or μ_1) plus random error
- Question: When there is no treatment effect is $Y = X$?

Two-sample Model: Formal Hypotheses

- Null Hypothesis: $H_0 : \mu_0 = \mu_1$
- Alternative Hypothesis: $H_1 : \mu_0 \neq \mu_1$
- Question:
  - Why is the null hypothesis not $\mu_0 \neq \mu_1$?
  - and the alternative hypothesis not $(\mu_0 = \mu_1)$?
- Assumptions:
  - Previous: In each population, the observed level is the true level plus noise
  - Additional: The random errors ϵ and ϵ' are normally distributed with mean 0 and common variance σ^2

Two-sample Model: Experimental Design

- Drawn random sample of size n from Untreated population
- This will yield n observations $X_1, \ldots, X_n$, where

$$X_i = \mu_0 + \epsilon_i$$

- Random sample of size n from Treated population
- This will yield n observations $Y_1, \ldots, Y_n$, where

$$Y_i = \mu_1 + \epsilon_i'$$

Two-sample Model: Assumptions

- The random errors $\epsilon_1, \ldots, \epsilon_n$ are mutually independent
- The random errors $\epsilon'_1, \ldots, \epsilon'_n$ are mutually independent
- $\epsilon_1, \ldots, \epsilon_n$ are normally distributed with mean 0 and variance σ^2
- $\epsilon'_1, \ldots, \epsilon'_n$ are normally distributed with mean 0 and variance σ^2
- The (two-sample) t-test is a commonly method for testing this hypothesis under the given set of assumptions (normal distribution with common variance)

Quick Note: Conservative versus Anti-conservative; Robustness

- The properties of the decision rule will depend on these underlying assumptions
- They may be greatly sensitive to these assumptions
- The type I error of a decision procedure we hope to achieve is called the nominal level
- Example: If we claim that the nominal level of our decision is 0.05, then we are asserting that the probability of committing a false-positive is at most 0.05.
- If the *actual* type I error rate exceeds the nominal level the test is said to be anti-conservative
- If the *actual* type I error rate is less than the nominal level the test is said to be conservative
- A decision rule that is not sensitive to the underlying assumptions, with respect to type I error control, is said to be robust

Designing the Experiment

- The sample size to achieve the desired power at a given type I error rate depends on the effect size
- Given everything else fixed, a larger effect size requires a smaller size to achieve a power at a given type I error rate
- The effect size for the two-sample t-test is defined as

$$\Delta = \frac{|\mu_0 - \mu_1|}{\sigma}$$

- The numerator $|\mu_0 - \mu_1|$ is the difference (in absolute value) of the means
- The size of this difference (how large it is) is in relation to (scaled by) the standard deviation

Sample Size Formula

- The sample size formula the two-sample t-test is

$$n = 2\frac{(Z_{1-\alpha} + Z_{1-\beta})^2}{\Delta^2}$$

- Here $Z_{1-\alpha}$ denote the right α tail of a normal distribution
- Let's forget most of the technical details
- Just observe that the sample size decreases as the effect size become larger. Why?
- Many other sample size formulas look very similar

Our Example: The Unknown Truth

- The true values of the unknown parameters:
  - $\mu_0 = 0$
  - $\mu_1 = 2$
  - $\sigma = 5$
- The effect size is

$$\Delta = \frac{|0 - 2|}{5} = 0.4$$

Forget about the Design

- What is the power if we use 3 units per group

```
##
##      Two-sample t test power calculation
##
##               n = 3
##           delta = 2
##              sd = 5
##       sig.level = 0.05
##           power = 0.05784303
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Forget about the Design

- What is the power if we use 6 units per group

```
##
##      Two-sample t test power calculation
##
##              n = 6
##          delta = 2
##             sd = 5
##      sig.level = 0.05
##          power = 0.09156966
##    alternative = two.sided
##
## NOTE: n is number in *each* group
```

Forget about the Design

- What is the power if we use 12 units per group

```
##
##      Two-sample t test power calculation
##
##              n = 12
##          delta = 2
##             sd = 5
##      sig.level = 0.05
##          power = 0.1532882
##    alternative = two.sided
##
## NOTE: n is number in *each* group
```

Now Use Experimental Design

- The required sample size, per group, to detect an effect size of

$$\Delta = \frac{|0-2|}{5} = 0.4$$

with a power of 0.8, at the 0.05 level is n =100 *per* group

```
##
##      Two-sample t test power calculation
##
##              n = 99.08057
##          delta = 2
##             sd = 5
##      sig.level = 0.05
##          power = 0.8
##    alternative = two.sided
##
## NOTE: n is number in *each* group
```

- If a smaller sample size is used, the study will be underpowered
- What is the caveat with using a larger sample size?
- Note: These observations are based on the given assumptions, effect size and type I and II errors

How to check the type I Error and Power

- Simulation provides a powerful framework for understanding the properties of the decision rule
- In the case of the two-sample t-test this works as follows
  1. Draw a random sample of size n from a normal distribution with mean μ_0 and standard deviation σ
  2. Draw a random sample of size n from a normal distribution with mean μ_1 and standard deviation σ
  3. Apply the two-sample test to the two data samples and record the P-value
- Now repeat the last three steps a large number of times
- The distribution of these simulated P-values should be similar to the true distribution of the P-value

Simulation Example

- Set parameters

```
set.seed(4141)
n = 6
mu0 = 0
mu1 = 2
sigma = 5
```

- Simulate data

```
x0 = rnorm(n, mu0, sigma)
x1 = rnorm(n, mu1, sigma)
x0
```

```
## [1] -2.1071177 -0.2402046  2.6668539 -4.4884699  2.6865668  5.1362518
```

```
x1
```

```
## [1]  6.0170556 -4.3949286 -1.4848887 -3.5189476 -8.7897573 -0.4961073
```

- Carry out t-test

```
t.test(x0, x1)
```

```
##
##  Welch Two Sample t-test
##
## data:  x0 and x1
## t = 1.0984, df = 9.1035, p-value = 0.3002
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.872410  8.312895
## sample estimates:
## mean of x mean of y
##  0.608980 -2.111262
```

Simulation: Important Notes

- Data are generated under the truth
- Parameters and distributions are set by you
- A simulated experiment is to mimic a hypothetical, but real, experiment
- The truth is not known in the context of a real experiment
- IMPORTANT: The decision rule step has to remain *blinded* to this truth
- Computing Exercise: Evaluate the type I error and power for the two-sample example using simulation and formula

Stat 101 Example: One-sided or Two-sided Test?

- Suppose that company XYZ Dairies sells milk in glass bottles
- The company claims that the net content of each bottle is 1 gallon
- Mr. Smith, owner of the ABC Supermarket, suspects he, and ultimately his customers, are being swindled by XYZ
- Let μ denote the mean net content (in gallons) of the *population* of XYZ Dairies milk bottles
- The company claims $\mu = 1$
- Mr. Smith hypothesizes that $\mu < 1$

Stat 101 Example (null vs alternative)

- Mr. Smith has to give benefit of the doubt to company XYZ's claim (i.e., $\mu = 1$)
- The purpose of the experiment is to ascertain if there is sufficient evidence to the contrary (i.e., show $\mu \neq 1$)
- The null hypothesis is formulated as $H_0 : \mu = 0$
- The alternative is formulated as $H_1 : \mu \neq 0$
- Mr. Smith has no interest in gathering evidence for showing that XYZ overfills its bottles (i.e., $\mu > 1$
- In this case, a one-sided hypothesis would be appropriate

Statistical versus Clinical/Biological Significance

- Hypothesis testing is carried out to investigate *statistical* and not *biological* significance
- It is the responsibility of the investigator to pose a biologically relevant hypothesis.
- It is also the responsibility of the investigator to ensure that a statistically significant finding is biologically plausible/realistic
- Statistical significance does not necessarily imply biological significance or vice versa

Biologically but not Statistically Significant

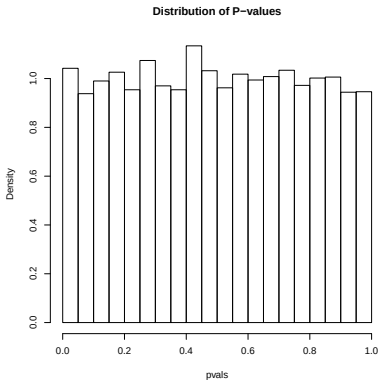
```
set.seed(1122333)
x0 = rnorm(3, 1, 1)
x1 = rnorm(3, 2, 1)
x0

## [1] -0.25824011  0.02820527  2.20878939

x1

## [1] 1.5462733 0.6578732 3.1782064

t.test(x0, x1)

##
##  Welch Two Sample t-test
##
## data:  x0 and x1
## t = -1.0572, df = 3.9884, p-value = 0.3502
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.117361  1.848295
## sample estimates:
## mean of x mean of y
## 0.6595849 1.7941176
```

Statistically but not Biologically Significant

```
x0 = c(3.0001, 3.0002, 3.0003, 3.0004, 3.0005)
x1 = c(3.0006, 3.0007, 3.0008, 3.0009, 3.001)
x0

## [1] 3.0001 3.0002 3.0003 3.0004 3.0005

x1

## [1] 3.0006 3.0007 3.0008 3.0009 3.0010

t.test(x0, x1)

##
##  Welch Two Sample t-test
##
## data:  x0 and x1
## t = -5, df = 8, p-value = 0.001053
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.0007306004 -0.0002693996
## sample estimates:
## mean of x mean of y
##    3.0003    3.0008
```

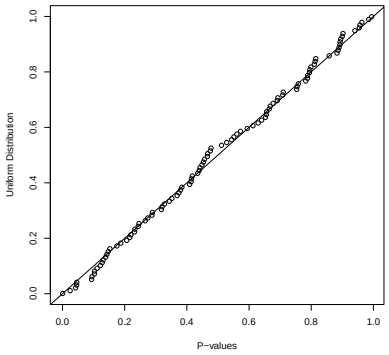
Distribution of P-values under H_0

- Under the null hypothesis, the distribution of the P-values is uniform
- If you repeat the experiment many times under the null hypothesis (e.g., no differential expression), the distribution of the P-values will look like this



Quantile-Quantile Plot

- An important tool to assess type I error control is the Quantile-Quantile Plot (aka QQ-Plot)
- The plot should look like this under H_0



Quantile-Quantile Plot: Deviation

- A deviation in the QQ-Plot indicates that there may be evidence to reject H_0
- Or that the decision rule is not accounting for type I error: INFLATION!!