Introduction to Biostatistics, Lecture 1

Xing Qiu

Department of Biostatistics and Computational Biology
University of Rochester
Office: MRBX-G.11311N
Email: xqiu@bst.rochester.edu

November 12, 2007

Outline

The Roles of Biostatistics in Medical Research

- Point Estimation
  - Calculating the mean, variance of a given sample, ...
  - Making nice tables, plots to visualize a given data.
- Hypothesis Testing
  - Given a *null hypothesis*, such as "no difference between the control and treatment group", how likely is our observed differences caused by chance only?
  - Given some reasonable conditions, such as "we know there is a 1.6 units difference between the control and treatment group", how many observations (patients) do we need in order to reject the *null hypothesis*? This question is answered by a power calculation.

The Roles of Biostatistics in Medical Research

- Point Estimation
  - Calculating the mean, variance of a given sample, . . .
  - Making nice tables, plots to visualize a given data.
- Hypothesis Testing
  - Given a *null hypothesis*, such as "no difference between the control and treatment group", how likely is our observed differences caused by chance only?
  - Given some reasonable conditions, such as "we know there is a 1.6 units difference between the control and treatment group", how many observations (patients) do we need in order to reject the *null hypothesis*? This question is answered by a power calculation.

The Roles of Biostatistics in Medical Research

- Point Estimation
  - Calculating the mean, variance of a given sample, . . .
  - Making nice tables, plots to visualize a given data.
- Hypothesis Testing
  - Given a *null hypothesis*, such as “no difference between the control and treatment group”, how likely is our observed differences caused by chance only?
  - Given some reasonable conditions, such as “we know there is a 1.6 units difference between the control and treatment group”, how many observations (patients) do we need in order to reject the *null hypothesis*? This question is answered by a power calculation.

The Roles of Biostatistics in Medical Research

- Point Estimation
  - Calculating the mean, variance of a given sample, . . .
  - Making nice tables, plots to visualize a given data.
- Hypothesis Testing
  - Given a *null hypothesis*, such as "no difference between the control and treatment group", how likely is our observed differences caused by chance only?
  - Given some reasonable conditions, such as "we know there is a 1.6 units difference between the control and treatment group", how many observations (patients) do we need in order to reject the *null hypothesis*? This question is answered by a power calculation.

The Roles of Biostatistics in Medical Research

- Point Estimation
  - Calculating the mean, variance of a given sample, . . .
  - Making nice tables, plots to visualize a given data.
- Hypothesis Testing
  - Given a *null hypothesis*, such as “no difference between the control and treatment group”, how likely is our observed differences caused by chance only?
  - Given some reasonable conditions, such as “we know there is a 1.6 units difference between the control and treatment group”, how many observations (patients) do we need in order to reject the *null hypothesis*? This question is answered by a power calculation.

The Roles of Biostatistics in Medical Research

- Point Estimation
  - Calculating the mean, variance of a given sample, . . .
  - Making nice tables, plots to visualize a given data.
- Hypothesis Testing
  - Given a *null hypothesis*, such as "no difference between the control and treatment group", how likely is our observed differences caused by chance only?
  - Given some reasonable conditions, such as "we know there is a 1.6 units difference between the control and treatment group", how many observations (patients) do we need in order to reject the *null hypothesis*? This question is answered by a power calculation.

The Roles of Biostatistics in Medical Research

- Modeling
  - Linear Regression, causal analysis, making predictions etc.

The Roles of Biostatistics in Medical Research

- Modeling
  - Linear Regression, causal analysis, making predictions etc.

The Population

Definition In statistics, a Population is the set of all persons, objects or events that share some characteristics of interest. study.

Examples Say you want to study the correlation between blood pressure and smoking.

- all persons that have smoked
- all persons with hypertension

Remark The population you really care about is all (potential) smokers in the world, now and in the future. Of course it is not feasible to study them one by one, so you decide to study, say 40 of them. These forty constitute a sample.

The Population

Definition In statistics, a Population is the set of all persons, objects or events that share some characteristics of interest. study.

Examples Say you want to study the correlation between blood pressure and smoking.

- all persons that have smoked
- all persons with hypertension

Remark The population you really care about is all (potential) smokers in the world, now and in the future. Of course it is not feasible to study them one by one, so you decide to study, say 40 of them. These forty constitute a sample.

The Population

Definition In statistics, a Population is the set of all persons, objects or events that share some characteristics of interest. study.

Examples Say you want to study the correlation between blood pressure and smoking.

- all persons that have smoked
- all persons with hypertension

Remark The population you really care about is all (potential) smokers in the world, now and in the future. Of course it is not feasible to study them one by one, so you decide to study, say 40 of them. These forty constitute a sample.

The Population

Definition In statistics, a Population is the set of all persons, objects or events that share some characteristics of interest. study.

Examples Say you want to study the correlation between blood pressure and smoking.

- all persons that have smoked
- all persons with hypertension

Remark The population you really care about is all (potential) smokers in the world, now and in the future. Of course it is not feasible to study them one by one, so you decide to study, say 40 of them. These forty constitute a sample.

The Population

Definition In statistics, a Population is the set of all persons, objects or events that share some characteristics of interest. study.

Examples Say you want to study the correlation between blood pressure and smoking.

- all persons that have smoked
- all persons with hypertension

Remark The population you really care about is all (potential) smokers in the world, now and in the future. Of course it is not feasible to study them one by one, so you decide to study, say 40 of them. These forty constitute a sample.

The Sample

- In statistics, a sample is a (usually tiny) fraction of the population that is choosen (observed) in a study.
- The usefulness of a sample depends on one thing: how well the sample represents the population. It in turn depends on two things: the sample size (how many observations) and the biasness.
- Of course, the larger the sample size the better. But there is always a budget constraint. That's why we need a good power calculation.
- As for the biasness: randomizing, diversity, common sense. Believe or not, this step is actually not part of statistics!

The Sample

- In statistics, a sample is a (usually tiny) fraction of the population that is choosen (observed) in a study.
- The usefulness of a sample depends on one thing: how well the sample represents the population. It in turn depends on two things: the sample size (how many observations) and the biasness.
- Of course, the larger the sample size the better. But there is always a budget constraint. That's why we need a good power calculation.
- As for the biasness: randomizing, diversity, common sense. Believe or not, this step is actually not part of statistics!

The Sample

- In statistics, a sample is a (usually tiny) fraction of the population that is choosen (observed) in a study.
- The usefulness of a sample depends on one thing: how well the sample represents the population. It in turn depends on two things: the sample size (how many observations) and the biasness.
- Of course, the larger the sample size the better. But there is always a budget constraint. That's why we need a good power calculation.
- As for the biasness: randomizing, diversity, common sense. Believe or not, this step is actually not part of statistics!

The Sample

- In statistics, a sample is a (usually tiny) fraction of the population that is choosen (observed) in a study.
- The usefulness of a sample depends on one thing: how well the sample represents the population. It in turn depends on two things: the sample size (how many observations) and the biasness.
- Of course, the larger the sample size the better. But there is always a budget constraint. That's why we need a good power calculation.
- As for the biasness: randomizing, diversity, common sense. Believe or not, this step is actually not part of statistics!

Types of Data

- Continuous data, e.g., blood pressure, weight
- Discrete data, e.g., number of patients in a hospital
- Binary data, a special case of discrete data, e.g., smoker/non-smoker
- Nominal data, not numerical, e.g., gender of patients
- Ordinal data, a special case of nominal data, with a natural order

Types of Data

- Continuous data, e.g., blood pressure, weight
- Discrete data, e.g., number of patients in a hospital
- Binary data, a special case of discrete data, e.g., smoker/non-smoker
- Nominal data, not numerical, e.g., gender of patients
- Ordinal data, a special case of nominal data, with a natural order

Types of Data

- Continuous data, e.g., blood pressure, weight
- Discrete data, e.g., number of patients in a hospital
- Binary data, a special case of discrete data, e.g., smoker/non-smoker
- Nominal data, not numerical, e.g., gender of patients
- Ordinal data, a special case of nominal data, with a natural order

Types of Data

- Continuous data, e.g., blood pressure, weight
- Discrete data, e.g., number of patients in a hospital
- Binary data, a special case of discrete data, e.g., smoker/non-smoker
- Nominal data, not numerical, e.g., gender of patients
- Ordinal data, a special case of nominal data, with a natural order

Types of Data

- Continuous data, e.g., blood pressure, weight
- Discrete data, e.g., number of patients in a hospital
- Binary data, a special case of discrete data, e.g., smoker/non-smoker
- Nominal data, not numerical, e.g., gender of patients
- Ordinal data, a special case of nominal data, with a natural order

Types of Data (2)

The classification of data by their types are not strict. Example:

- Expensive/Economic: Nominal
- Label Expensive as 1, Economic as 0: binary
- Very expensive (VE)/expensive (E)/economic (C)/very economic (VC): Nominal again, not binary
- VE > E > C > VC, ordinal
- Dollar amount: 10, 5, 2, 1 (Unit = \$1000), discrete
- Exact dollar amount: 10.321, 5.640, etc, (more or less) continuous

Types of Data (2)

The classification of data by their types are not strict. Example:

- Expensive/Economic: Nominal
- Label Expensive as 1, Economic as 0: binary
- Very expensive (VE)/expensive (E)/economic (C)/very economic (VC): Nominal again, not binary
- VE > E > C > VC, ordinal
- Dollar amount: 10, 5, 2, 1 (Unit = $1000), discrete
- Exact dollar amount: 10.321, 5.640, etc, (more or less) continuous

Types of Data (2)

The classification of data by their types are not strict. Example:

- Expensive/Economic: Nominal
- Label Expensive as 1, Economic as 0: binary
- Very expensive (VE)/expensive (E)/economic (C)/very economic (VC): Nominal again, not binary
- VE > E > C > VC, ordinal
- Dollar amount: 10, 5, 2, 1 (Unit = $1000), discrete
- Exact dollar amount: 10.321, 5.640, etc, (more or less) continuous

Types of Data (2)

The classification of data by their types are not strict. Example:

- Expensive/Economic: Nominal
- Label Expensive as 1, Economic as 0: binary
- Very expensive (VE)/expensive (E)/economic (C)/very economic (VC): Nominal again, not binary
- VE > E > C > VC, ordinal
- Dollar amount: 10, 5, 2, 1 (Unit = $1000), discrete
- Exact dollar amount: 10.321, 5.640, etc, (more or less) continuous

Types of Data (2)

The classification of data by their types are not strict. Example:

- Expensive/Economic: Nominal
- Label Expensive as 1, Economic as 0: binary
- Very expensive (VE)/expensive (E)/economic (C)/very economic (VC): Nominal again, not binary
- VE > E > C > VC, ordinal
- Dollar amount: 10, 5, 2, 1 (Unit = $1000), discrete
- Exact dollar amount: 10.321, 5.640, etc, (more or less) continuous

Types of Data (2)

The classification of data by their types are not strict. Example:

- Expensive/Economic: Nominal
- Label Expensive as 1, Economic as 0: binary
- Very expensive (VE)/expensive (E)/economic (C)/very economic (VC): Nominal again, not binary
- VE > E > C > VC, ordinal
- Dollar amount: 10, 5, 2, 1 (Unit = $1000), discrete
- Exact dollar amount: 10.321, 5.640, etc, (more or less) continuous

The Arithmetic Mean

This is the most frequently used measure of "center" of a sample. Def and example: ...
Caution:

- it doesn't apply to nominal data.
- sensitive to an "outlier" (def/reason of an outlier)
- It is very important to know that sample mean is only an approximation to the population mean we are looking for.

The Median

- Suppose you have a series of n values: X_1, X_2, X_n
- The median is defined as
- If n is odd: the $(n+1)/2$th largest value
- If n is even: the average of the $n/2$th and $(n/2+1)$th largest values
- Equal numbers of values on both sides of the median.

The Median

- Suppose you have a series of n values: X_1, X_2, X_n
- The median is defined as
- If n is odd: the $(n+1)/2$th largest value
- If n is even: the average of the $n/2$th and $(n/2+1)$th largest values
- Equal numbers of values on both sides of the median.

The Median

- Suppose you have a series of n values: X_1, X_2, X_n
- The median is defined as
- If n is odd: the $(n+1)/2$th largest value
- If n is even: the average of the $n/2$th and $(n/2+1)$th largest values
- Equal numbers of values on both sides of the median.

The Median

- Suppose you have a series of n values: X_1, X_2, X_n
- The median is defined as
- If n is odd: the $(n+1)/2$th largest value
- If n is even: the average of the $n/2$th and $(n/2+1)$th largest values
- Equal numbers of values on both sides of the median.

The Median

- Suppose you have a series of n values: X_1, X_2, X_n
- The median is defined as
- If n is odd: the $(n+1)/2$th largest value
- If n is even: the average of the $n/2$th and $(n/2+1)$th largest values
- Equal numbers of values on both sides of the median.

Comparison of the Mean and the Median

- Sample: $X_1 = 3.1, X_2 = 2.2, X_3 = 6.7, X_4 = 1.8$
- Mean: $\bar{X} = (X_1 + X_2 + X_3 + X_4)/4 = 3.45$
- Median: the average of 3.1 and 2.2, which is 2.65
- Change X_1 to 31 (an outlier)
- Mean: 10.425; Median: 4.45
- Is it a strength or a weakness? Relation to the normal distribution

Comparison of the Mean and the Median

- Sample: $X_1 = 3.1, X_2 = 2.2, X_3 = 6.7, X_4 = 1.8$
- Mean: $\bar{X} = (X_1 + X_2 + X_3 + X_4)/4 = 3.45$
- Median: the average of 3.1 and 2.2, which is 2.65
- Change X_1 to 31 (an outlier)
- Mean: 10.425; Median: 4.45
- Is it a strength or a weakness? Relation to the normal distribution

Comparison of the Mean and the Median

- Sample: $X_1 = 3.1, X_2 = 2.2, X_3 = 6.7, X_4 = 1.8$
- Mean: $\bar{X} = (X_1 + X_2 + X_3 + X_4)/4 = 3.45$
- Median: the average of 3.1 and 2.2, which is 2.65
- Change X_1 to 31 (an outlier)
- Mean: 10.425; Median: 4.45
- Is it a strength or a weakness? Relation to the normal distribution

Comparison of the Mean and the Median

- Sample: $X_1 = 3.1, X_2 = 2.2, X_3 = 6.7, X_4 = 1.8$
- Mean: $\bar{X} = (X_1 + X_2 + X_3 + X_4)/4 = 3.45$
- Median: the average of 3.1 and 2.2, which is 2.65
- Change X_1 to 31 (an outlier)
- Mean: 10.425; Median: 4.45
- Is it a strength or a weakness? Relation to the normal distribution

Comparison of the Mean and the Median

- Sample: $X_1 = 3.1, X_2 = 2.2, X_3 = 6.7, X_4 = 1.8$
- Mean: $\bar{X} = (X_1 + X_2 + X_3 + X_4)/4 = 3.45$
- Median: the average of 3.1 and 2.2, which is 2.65
- Change X_1 to 31 (an outlier)
- Mean: 10.425; Median: 4.45
- Is it a strength or a weakness? Relation to the normal distribution

Comparison of the Mean and the Median

- Sample: $X_1 = 3.1, X_2 = 2.2, X_3 = 6.7, X_4 = 1.8$
- Mean: $\bar{X} = (X_1 + X_2 + X_3 + X_4)/4 = 3.45$
- Median: the average of 3.1 and 2.2, which is 2.65
- Change X_1 to 31 (an outlier)
- Mean: 10.425; Median: 4.45
- Is it a strength or a weakness? Relation to the normal distribution

Et Cetera

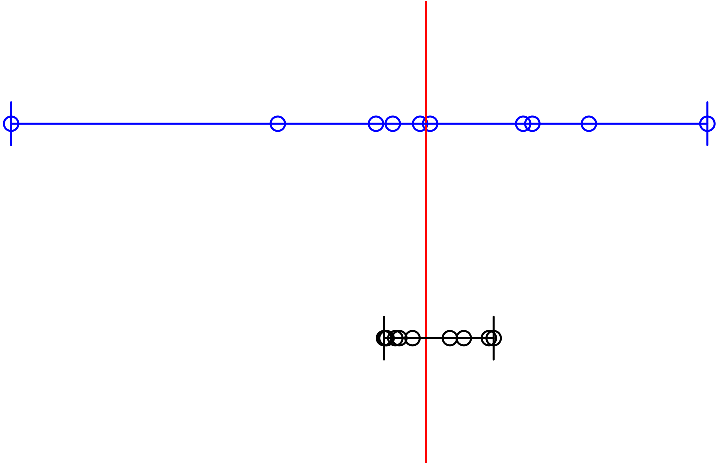
- Modes: most frequently observed point
- Geometric Mean: log transformation ==> mean ==> antilogarithm

Et Cetera

- Modes: most frequently observed point
- Geometric Mean: log transformation ==> mean ==> antilogarithm

Introduction

Two samples with the same center but different variation



The Range

- Sample: $X_1 = 3.1, X_2 = 2.2, X_3 = 6.7, X_4 = 1.8$
- Minimum: $X_4 = 1.8$
- Maximum: $X_3 = 6.7$
- Range: 6.7-1.8 = 4.9
- Sometimes a range is denoted as (1.8, 6.7)
- Disadvantage: very sensitive to the extreme values, it also depends on the sample size

The Range

- Sample: $X_1 = 3.1, X_2 = 2.2, X_3 = 6.7, X_4 = 1.8$
- Minimum: $X_4 = 1.8$
- Maximum: $X_3 = 6.7$
- Range: 6.7-1.8 = 4.9
- Sometimes a range is denoted as (1.8, 6.7)
- Disadvantage: very sensitive to the extreme values, it also depends on the sample size

The Range

- Sample: $X_1 = 3.1, X_2 = 2.2, X_3 = 6.7, X_4 = 1.8$
- Minimum: $X_4 = 1.8$
- Maximum: $X_3 = 6.7$
- Range: 6.7-1.8 = 4.9
- Sometimes a range is denoted as (1.8, 6.7)
- Disadvantage: very sensitive to the extreme values, it also depends on the sample size

The Range

- Sample: $X_1 = 3.1, X_2 = 2.2, X_3 = 6.7, X_4 = 1.8$
- Minimum: $X_4 = 1.8$
- Maximum: $X_3 = 6.7$
- Range: 6.7-1.8 = 4.9
- Sometimes a range is denoted as (1.8, 6.7)
- Disadvantage: very sensitive to the extreme values, it also depends on the sample size

The Range

- Sample: $X_1 = 3.1, X_2 = 2.2, X_3 = 6.7, X_4 = 1.8$
- Minimum: $X_4 = 1.8$
- Maximum: $X_3 = 6.7$
- Range: 6.7-1.8 = 4.9
- Sometimes a range is denoted as (1.8, 6.7)
- Disadvantage: very sensitive to the extreme values, it also depends on the sample size

The Range

- Sample: $X_1 = 3.1, X_2 = 2.2, X_3 = 6.7, X_4 = 1.8$
- Minimum: $X_4 = 1.8$
- Maximum: $X_3 = 6.7$
- Range: 6.7-1.8 = 4.9
- Sometimes a range is denoted as (1.8, 6.7)
- Disadvantage: very sensitive to the extreme values, it also depends on the sample size

Quantiles

- Roughly, the pth quantile(percentile), denoted as V_p, is the value such that approximately p percent of the sample observations are less than or equal to V_p.
- See page 18 for a full definition
- Median is the 50th quantile.
- Quartiles: 25th, 50th, and 75th quantiles.

Quantiles

- Roughly, the pth quantile(percentile), denoted as V_p, is the value such that approximately p percent of the sample observations are less than or equal to V_p.
- See page 18 for a full definition
- Median is the 50th quantile.
- Quartiles: 25th, 50th, and 75th quantiles.

Quantiles

- Roughly, the pth quantile(percentile), denoted as V_p, is the value such that approximately p percent of the sample observations are less than or equal to V_p.
- See page 18 for a full definition
- Median is the 50th quantile.
- Quartiles: 25th, 50th, and 75th quantiles.

Quantiles

- Roughly, the *p*th quantile(percentile), denoted as V_p, is the value such that approximately p percent of the sample observations are less than or equal to V_p.
- See page 18 for a full definition
- Median is the 50th quantile.
- Quartiles: 25th, 50th, and 75th quantiles.

The variance

- Page 19-21 is an excellent guide.
- mean signed deviation which doesn't work
- mean deviation
- sample variance

The variance

- Page 19-21 is an excellent guide.
- mean signed deviation which doesn't work
- mean deviation
- sample variance

The variance

- Page 19-21 is an excellent guide.
- mean signed deviation which doesn't work
- mean deviation
- sample variance

The variance

- Page 19-21 is an excellent guide.
- mean signed deviation which doesn't work
- mean deviation
- sample variance

Why $N-1$?

- population variance $\frac{\sum_{i=1}^{N}(X_i-\mu)}{N}$, when N is really large
- By using $\bar{X}$ instead of μ, one underestimate the variance because $\bar{X}$ is not a perfect estimate of μ.
- lost one degree of freedom
- The difference between the sample variance and the population variance is very important

STD and Coefficient of Variation

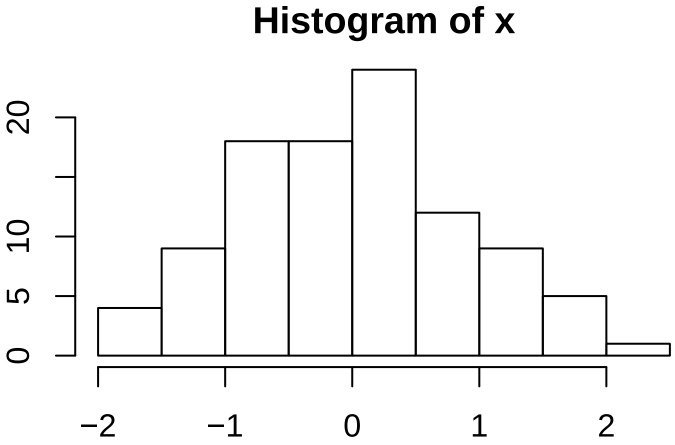
- Standard Deviation (STD): $\sqrt{}(Variance)$
- Coefficient of Variation: $\frac{STD}{Mean}$

Graphic Methods

Page 28-33

- Bar Plot/Histogram
- Stem-and-Leaf Plots
- Box plots

Histogram



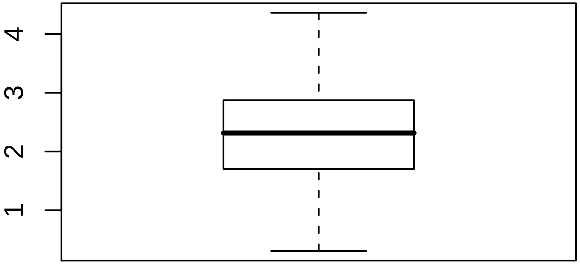
Stem-and-Leaf Plot

This is more like a table rather than a plot:

-0	976
0	58
1	45

Box Plot

The famous *Box and whiskers plot*. min, Q1, median, Q3 and max.



Next lecture

- Chapter 2, Probability
- Readings Page 33-37, case studies

References

- Rosner, B. (2006) Fundamentals of Biostatistics
- Motulsky, H. (1995) Intuitive Biostatistics