Outline

Contents

1 Introduction

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 (Cont.)

- Modeling
  - Linear Regression, causal analysis, making predictions etc.
- Other statistical procedures. Quality assurance analysis, outlier detection, cluster analysis, dimension reduction, decision tree, and many, many more.

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!

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

2 Descriptive Statistics

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.

Comparison of the Mean and the Median

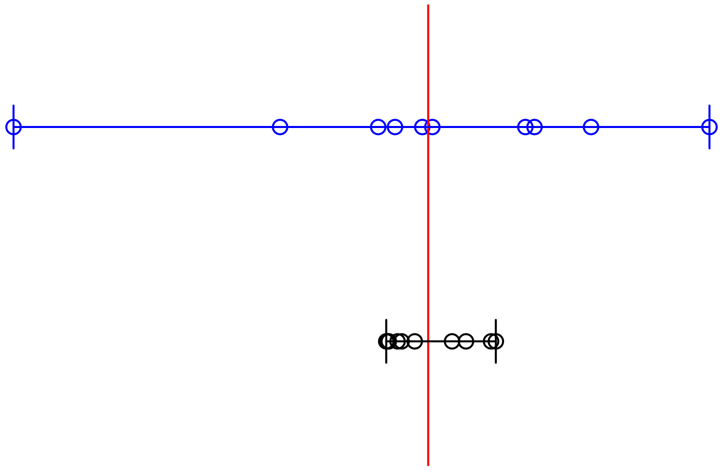
- Sample: $X_1 = 3.1, X_2 = 2.2, X_3 = 6.7, X_4 = 1.8$
- Mean: $\bar{X} = (X_1 + X_2 + X_3 + X_4)/4 = 3.45$
- Median: the average of 3.1 and 2.2, which is 2.65
- Change X_1 to 31 (an outlier)
- Mean: 10.425; Median: 4.45
- Is it a strength or a weakness? Relation to the normal distribution

Et Cetera

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

Measure of variation

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

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.

The variance

- sample variance is defined as
$$\hat{\sigma}^2 := \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}. \tag{1}$$
- Sample standard deviation: squared root of $\hat{\sigma}^2$.

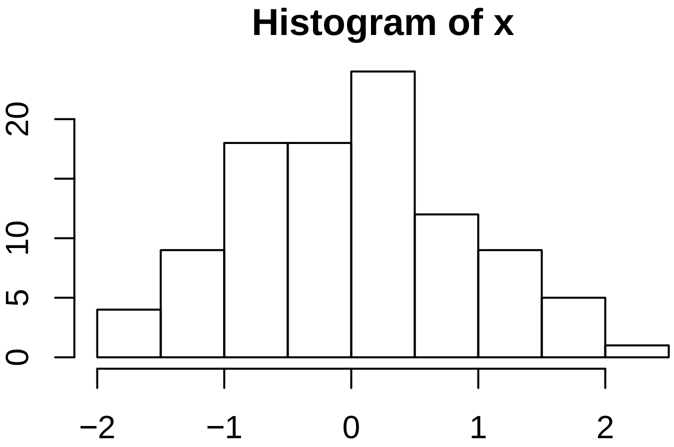
Why $n-1$?

- population variance is $\frac{\sum_{i=1}^n (X_i - \mu)^2}{n}$, 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. We lose one degree of freedom in this step.
- The difference between the sample variance and the population variance.

Graphic Methods

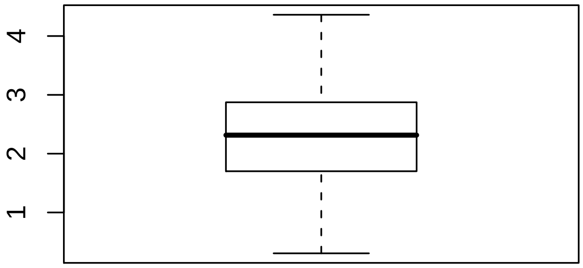
- Bar Plot/Histogram
- Box plots

Histogram



Box Plot

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



3 Probability distributions

Probability distributions

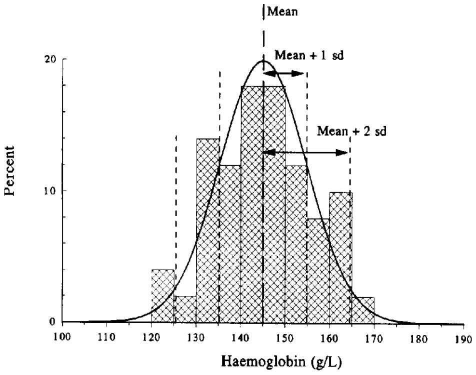
- Standard Bernoulli distribution (a fair coin). Explain the “coding” event.
- The discrete density function: $f(0)$ and $f(1)$.
- General Bernoulli distribution: unfair coin. One parameter: success rate.
- Binomial distribution and other discrete distributions.

Continuous distributions

- Normal distribution. Two parameters: mean and variance (sd).
- Non-normal distributions – thick tail, asymmetry, etc.
- Non-normal distributions may have more than two parameters. They do not have to be symmetric (mean $\neq$ median).
- Large sample approximation (central limit theorem in a nutshell).

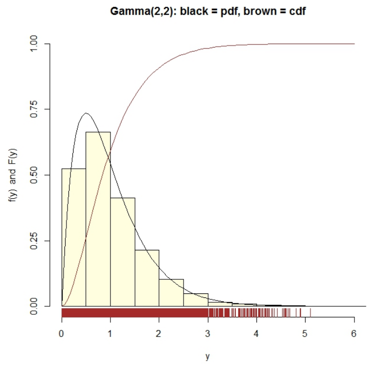
Normal density and histogram

The bell curve (normal distribution) $N(\mu, \sigma^2)$. Explain the relationship between histograms (frequencies) and the idealized probability density function (pdf).



Normal density and histogram

An example of non-normal distribution ($\Gamma(2,2)$). Also explain the relationship between pdf and cumulative distribution function (cdf).



4 Point Estimation

Model-based estimation

- Definition: an estimator of a parameter θ, usually denoted as $\hat{\theta}$, is a quantity defined from the **observations** that *approximates* the unknown value of population θ.

- $\hat{\theta}$ usually (if not always) differs from the true value of θ (This is a consequence of the sampling error).
- Heuristically speaking, the smaller this difference is, the better this estimator is. How to measure this *difference* can get a little complicated, though.

Examples of estimator

- The Maximum likelihood estimator (MLE). Observing the given data is "most likely" if the true parameter is set to $\hat{\theta}^{\text{mle}}$.
- Bias-adjusted MLE. $n \to n-1$ in the denominator.
- Use moment estimator when MLE is hard to derive/compute.

Confidence intervals

- $\hat{\theta}$ is a *random variable* which is an approximation of θ. How accurate this approximation is? This is why we need confidence intervals for the estimated parameters.
- Heuristically, a 95% confidence interval for θ is an interval $I = [C_1, C_2]$ such that we can say: we are 95% confident that this interval includes the true value of the parameter θ.

Relation to standard error of the mean (SEM)

- Sometimes, you see a mean estimate reported as 3.12 ± 0.8.
- This $\pm s$ notation can mean two very different things: estimated standard deviation; or the standard error of the mean (SEM).
- SEM = $\frac{\sigma}{\sqrt{n}}$. Reason: $\sigma^2(\bar{X}) = \frac{\sigma^2(X_i)}{n}$.
- The normal approximation: $\bar{X} \pm (2 \cdot \text{SEM})$ is approximately 95% CI.

5 Hypothesis Testing

Why we need to test hypotheses

- Say we want to answer questions like: Is the new treatment better than the existing one?
- One approach is to estimate the mean response from the new treatment group and compare it to that from the old treatment group; if the first one is better then we declare the new drug is better.
- Drawback: Even if there is *no difference*, 1/2 of the time we will report new drug better due to *randomness*.

General concepts

- The above problem can be expressed in terms of the following hypotheses:

  H_0: The 2 treatments have identical performances.

  H_1: The new one has better performances.

  Here H_0 is the null hypothesis and H_1 is the alternative hypothesis.

- Note that in practice, most of time we test the following *two-sided* alternative hypothesis instead.

H_1: The 2 treatments have different performances.

Decisions

- A test may have 2 possible outcomes
  1. We accept H_0 (The 2 treatments have identical performances).
  2. We reject H_1 and accept H_1 (The 2 treatments have different performances).

Type I, II error, testing power

	Accept H_0	Reject H_0
H_0	True Negative	False Positive, type I error)
H_1	False Negative, type II error	True Positive)

Type I error and statistical power

- The probability of making a type I error is often denoted as α.
- It's commonly referred to as the *significant level* of the test.
- The probability of making a type II error is often denoted as β.
- The *statistical power* of the test is defined as $1-\beta$, which is the probability of rejecting H_0 if H_1 is true.

Two group comparison with unknown variance

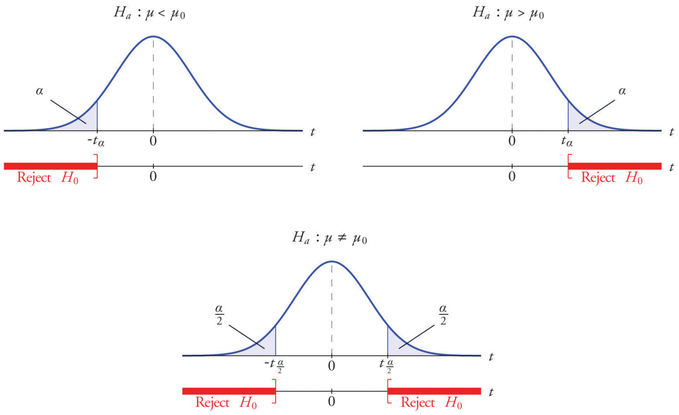
- A natural way to test H_0 versus H_1 is to use mean differences, $D = \bar{X}^A - barX^B$.
- Is $D = -12.88$ significant or not?
- Answer: We need to know the *variance*. For example, 12.88 pounds of mean body weight difference may be significant; 12.88 *grams* of difference may be trivial.
- A lot of hypothesis tests rely on some sort of "signal-to-noise ratio" statistic to make inference.

Two sample student t-statistic

- Definition
$$t = \frac{\bar{\mathbf{x}}^{(A)} - \bar{\mathbf{x}}^{(B)}}{\sqrt{s^2(\frac{1}{n_A} + \frac{1}{n_B})}}. \tag{2}$$
where s^2 is the estimated variance using data pooled from both groups.
- Under normal assumption, $t \sim T(n_A + n_B - 2)$, a Student's t-distribution with $n_A + n_B - 2$ degrees of freedom, which is very similar to a standard normal distribution when sample size is large. (Why?)

Inference

We can check the theoretical distribution of $T(n_A + n_B - 2)$ and reject H_0 if the value of t is very "unlikely".



Parametric versus nonparametric inference

- What if the underlying distribution is not normal?
- Inference based on t-distribution *may* be incorrect.
- It makes more sense to use a distribution-free statistical method, such as Wilcoxon rank-sum statistic, to perform the inference.

Wilcoxon rank-sum statistic

- The Wilcoxon rank-sum statistic is a nonparametric alternative to t-statistic:
$$W = \sum_{j=1}^{n_1} r(x_j^{(A)}) \tag{3}$$
- Where $r(x_j^{(A)})$ is the rank of $x_j^{(A)}$ in *both groups*.
- The key point: only ranking information are needed, so

1. p-value computed from this statistic is valid even when the distribution of the expression levels are not normal. [1]
2. outliers have less effect to inference.
3. Usually less powerful than its parametric counterpart.

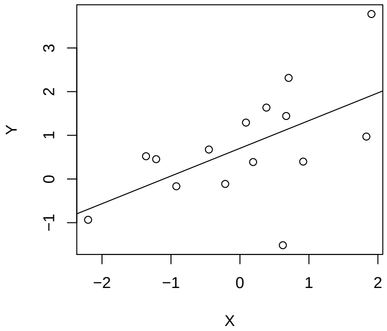
Other testing tools

- There are many, many specialized tests design for many, many different type of problems.
- Examples:
  1. Multiple group comparisons. One-way ANOVA F-test. Kruskal-Wallis test (nonparametric).
  2. Dependent data (as in pre/post analysis) group comparisons. Paired t-test; Friedman's test, etc.
  3. Binary data association (such as smoking and lung cancer). χ^2-test and Fisher's exact test.
  4. Linear regression. Goodness-of-fit F-test. t-test for linear association.
  5. Survival analysis. Log-rank test.
  6. Many, many more.
- If unsure, consult with a statistician for the most accurate and effective test for you data.

6 Linear Regression Analysis

Why linear regression

This figure shows us an example in which we want to find out the *linear relationship* between the covariate (X) and response (Y).



Ordinary linear regression model

- A simple linear model

$$Y = \beta_0 + \beta_1 X + \varepsilon. \tag{4}$$

- A more general linear model is given as follows

$$Y = \mathbf{X}\boldsymbol{\beta} + \varepsilon. \tag{5}$$

- Here X can be an $n \times p$-dimensional matrix. n is the sample size, and p is the number of covariates. Y is a vector of length n. ε is a vector of errors.
- In many applications, we assume that $\varepsilon_i \sim N(0, \sigma^2)$ and they are independent.

Inference of linear model

- We can estimate β_j by either maximum likelihood principle or minimizing the RSS. These two approaches are equivalent under the normal $i.i.d.$ assumption.
- More than often, we also want to test the *significance* of the linear association.
- The *global* approach: Goodness-of-fit F-test. Essentially it is a signal-to-noise ratio.
- Once Goodness-of-fit F-test declares that there is an overall significant linear association between covariates and Y, we can use t-test to determine *which* covariates contributed to this association.

Model selection

- If your goal is to build the best predictive model, using $p < 0.05$ cut for individual covariates is not the best way.
- Marginally important variables may not be significant in multiple regression!
- Stepwise model selection based on AIC/BIC; L^1-penalized regression for high-dimensional regressions.

Advanced Regression Models

- General linear model, which allows non-independent εs. Good for simple longitudinal analysis.
- Full-fledged linear mixed effect model. You can model both fixed effect (like the covariates in an ordinal regression) and random effects (subject effect, time effect, batch effect, etc) in one model.
- Non-normal responses (such as binary or categorical responses). Generalized linear model, possibly with random effect terms. One of the most popular GLM is logistic regression.
- Nonparametric regression, *a.k.a.* curve fitting.
- Other Specialized regression models, such as some time course regression models; spatial regression models, etc.

Other Types of Statistical Analyses

- Data transformation. z-transformation. High-throughput data normalization. Image registration.
- Dimension reduction. Principal component analysis/factor analysis.
- Model-based cluster analysis.

- Discriminant analysis. Decision making.
- Network inference methods.
- Many, many other procedures!

Bibliography

References

[1] Jean Dickinson Gibbons and Subhabrata Chakraborti, *Nonparametric statistical inference*, Marcel Dekker, Inc, 1992.