# Biostatistics (III)

Jianyong Sun
School of Mathematics and Statistics
Xi'an Jiaotong University

Apr., 2018

From previous frequency distribution, two obvious basic features: centrality and discreteness.

- centrality: the variable values tend to concentrate to a centre point
- discreteness: disperse around the center.

## Mean

- Arithmetic mean of the population: $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$
- Arithmetic mean of the samples: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Geometric mean: $G$: $G = \sqrt[n]{\prod_{i=1}^{n} x_i} \rightarrow$ for $x$ follows log-normal distribution.
- median: $M_d$: The median is the middle value of a set of values.
  - when $n$ is odd, median is the $\frac{n+1}{2}$-th value
  - wben $n$ is even, median is the average of the $\frac{n}{2}$-th and $(\frac{n}{2}+1)$-th values.
- mode:$M_o$: The mode of a set of data values is the value that appears most often.

## Statistical Concepts

Important properties of the arithmetic mean

- deviation from mean: difference between the observation values and the mean: the sum is zero

$$\sum (x - \bar{x}) = 0$$

- square of the difference between the observation to the mean (mean deviation sum of square) is minimal, i.e.

$$\sum (x - \bar{x})^2 < \sum (x - a)^2 \text{ for any } a \neq \bar{x}$$

Proof:

$$\begin{aligned}\sum (x-a)^2 &= \sum [(x - \bar{x}) + (\bar{x} - a)]^2 \\ &= \sum (x - \bar{x})^2 + 2\sum (x - \bar{x})(\bar{x} - a) + n(\bar{x} - a)^2\end{aligned}$$

## Variance or Variability

To measure discreteness.

- Range: $R = \max\{x_1, \cdots, x_n\} - \min\{x_1, \cdots, x_n\}$
- Sample variance: $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ where $n-1$ is the degree of freedom $df$, $s^2$ is the best estimate of $\sigma^2$.
- Population variance: $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$
- Standard deviation of the sample $s$
- Standard deviation of the population $\sigma$.
- Coefficient of variability, CV:

$$CV = \frac{s}{\bar{x}} \times 100\%$$

$s^2$ is a unbiased estimator, but

$$S^2 = \frac{\sum(x-\bar{x})^2}{n}$$

is a biased estimator of $\sigma^2$ since

$$\begin{aligned} E[S^2] &= E\left[\frac{1}{n}\sum(x_i-\bar{x})^2\right] = E\left[\frac{1}{n}\sum\left((x_i-\mu)-(\bar{x}-\mu)\right)^2\right] \\ &= E\left[\frac{1}{n}\sum(x_i-\mu)^2 - \frac{2}{n}(\bar{x}-\mu)\sum(x_i-\mu)+(\bar{x}-\mu)^2\right] \\ &= E\left[\frac{1}{n}\sum(x_i-\mu)^2\right] - E\left[(\bar{x}-\mu)^2\right] \\ &= \sigma^2 - E\left[(\bar{x}-\mu)^2\right] = (1-\frac{1}{n})\sigma^2 < \sigma^2 \end{aligned}$$

# Statistical Concepts

standard deviation of the sample $s$ is

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

standard deviation of the population $\sigma$ is

$$\sigma = \sqrt{\frac{\sum(x - \mu)^2}{N}}$$

The computation of the std.

$$\begin{aligned} \sum(x - \bar{x})^2 &= \sum(x^2 - 2x\bar{x} + \bar{x}^2) \\ &= \sum x^2 - \frac{(\sum x)^2}{n} \end{aligned}$$

## coefficient of variability, CV

when comparing two samples, it is not appropriate to describe their degrees of variability. To overcome, use coefficient of variability (CV).

- it is a relative value to the sample variable
- it has no units.
- it can be used to compare the variability of different samples.

# Probability Distribution

## Discrete random variable

- A random variable $x$ takes only discrete values from $\{x_i, i = 1, \cdots, n\}$
- A probability for each $x_i$: $P(x = x_i) = p_i, (i = 1, 2, \cdots, n)$

## Continuous random variable

- A random variable $x$ takes continuous values from $\Omega$.
- The probability of $x$ takes values in $[x_1, x_2]$ is

$$P(x_1 \le x \le x_2) = \int_{x_1}^{x_2} f(x) dx$$

where $f(x)$ is the probability density function (PDF).

## Law of large number

It is to describe the stability of a large number of random experiments.

## Theorem

*Let $m$ is the number of the appearance of event $A$ in $n$ independent random experiments, $p$ is the probability of the appearance of $A$, then for any positive number $\epsilon$, we have*

$$\lim_{n\to\infty} P\left\{\left|\frac{m}{n} - p\right| < \epsilon\right\} = 1$$

Khinchine theorem: to proof why arithmetic mean of the sample $\bar{x}$ can be used to infer the arithmetic mean of the population $\mu$

$$\lim_{n\to\infty} P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} x_i - \mu\right| < \epsilon\right\} = 1$$

Commonly-used Theoretical Distributions

## Binomial distribution

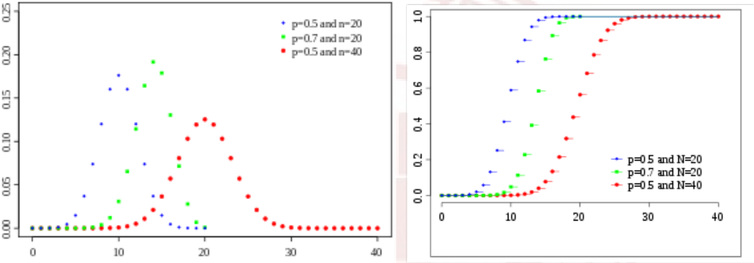
- $x$: the number of appearances of an event $A$ in $n$ random experiments
- Its probability mass function:

$$P(x) = C_n^x p^x (1-p)^{n-x} : B(n, p)$$

- its probability cumulative function:

$$F(x) = P(x \leq i) = \sum_{x=0}^{i} P(x)$$

# Probability Distribution




(a) PMF (b) CDF

$$\begin{aligned} \text{Mean} &= np \\ \text{Median} &= \lceil np \rceil \text{or } \lfloor np \rfloor \\ \text{Mode} &= \lceil (n+1)p \rceil \text{or } \lfloor (n+1)p \rfloor - 1 \\ \text{Variance} &= np(1-p) \end{aligned}$$

## Poisson distribution

- $x$: the number of appearances of an event $A$ in $n$ random experiments, but with small $p$ and large $n$
- Its probability mass function:

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

where $\lambda = np$

- its mean, variance, mode and median:

$$\lambda, \qquad \lambda, \qquad \lceil\lambda\rceil - 1 \text{ or } \lfloor\lambda\rfloor, \qquad \approx \lfloor\lambda + 1/3 - 0.02/\lambda\rfloor$$

## Normal distribution

- $\mathbf{x} \in \mathbb{R}^d$: continuous random variable
- Its probability density function:

$$f(\mathbf{x}) = \frac{|\Sigma|^{-1/2}}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\intercal \Sigma^{-1}(\mathbf{x}-\mu)\right) : \mathcal{N}(\mu, \Sigma)$$

- its mean, variance, mode and median:

$$\mu, \qquad \Sigma, \qquad \mu \qquad \mu$$

- Standard normal distribution:

$$u = \Sigma^{-1/2}(\mathbf{x}-\mu) \sim \mathcal{N}(0, \mathbf{I})$$

## The distribution of statistics

It concerns the relationship between samples and population.

- from population to samples:
  - sampling from the population, and check the differences between samples and population
  - study the distribution and statistics of the sampling
- from samples to population
  - from a sample or a series of samples to infer the population
  - statistical inference.

## Sampling Experiment

- The sampling procedure must obey the principle of randomness
- It is not possible to sample all individuals from the population
- Only sampling a small part: sampling with replacement

## Distribution of the sample mean

- multiple samplings – multiple means of samples $\bar{x}$ (random variable)
- The mean and variance of the sample mean $\bar{x}$

$$
\begin{aligned}
\mu_{\bar{x}} &= \frac{\sum f\bar{x}}{N^n} \\
\sigma_{\bar{x}}^2 &= \frac{1}{N^n}\left[\sum f\bar{x}^2 - \frac{\left(\sum f\bar{x}\right)^2}{N^n}\right]
\end{aligned}
$$

where $N$ is the size of population, $n$ is the sample size, and $f$ is the times of samplings of taking $\bar{x}$

The distribution of the sample mean

$$\mu_{\bar{x}} = \mu$$
$$\sigma_{\bar{x}} = \frac{\sigma^2}{n}$$

Further,

- If sampling from $\mathcal{N}(\mu, \sigma^2)$ then $\bar{x} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$
- If sampling not from normal, but with $\mu$ and $\sigma^2$, when the sampling size $n$ gets bigger, $\bar{x}$ approaches to $\mathcal{N}(\mu, \frac{\sigma^2}{n})$ more $\rightarrow$ central limit theorem
- $n \geq 30$: large sample, central limit theorem can be applied.

Consider an example: Let $N = 3$ for an approximate normal population, taking values $\{3, 4, 5\}$. Its $\mu = 4, \sigma^2 = 0.6667, \sigma = 0.8165$. Take $n = 2$ for sampling with replacement, we have a total of $N^n = 9$ samples.

| NO. | Samples | $\bar{x}$ | $s^2$ | s |
|---|---|---|---|---|
| 1 | 3, 3 | 3.0 | 0.0 | 0.0000 |
| 2 | 3, 4 | 3.5 | 0.5 | 0.7071 |
| 3 | 3, 5 | 4.0 | 2.0 | 1.4142 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 7 | 5, 3 | 4.0 | 2.0 | 1.4142 |
| 8 | 5, 4 | 4.5 | 0.5 | 0.7071 |
| 9 | 5, 5 | 5.0 | 0.0 | 0.0000 |
| $\Sigma$ | | 36 | 6.0 | 5.6568 |

The mean of the sample mean $\bar{x}$, i.e. $\mu_{\bar{x}} = \frac{36}{9} = 4$, and the mean of the sample variance $s^2$: $\mu_{s^2} = \frac{6}{9} = 0.6667 = \sigma^2$, but the mean of the sample standard deviation $\mu_s = \frac{5.6568}{9} = 0.6285 \neq \sigma$

# The distribution of statistics

For $N = 3$, $n = 4$, the total number of samples, $N^n = 81$

| $n=2$ | | | | $n=4$ | | | |
|---|---|---|---|---|---|---|---|
| $\bar{x}$ | times | $f\bar{x}$ | $f\bar{x}^2$ | $\bar{x}$ | times | $f\bar{x}$ | $f\bar{x}^2$ |
| 3.0 | 1 | 3 | 9.0 | 3.00 | 1 | 3 | 9.00 |
| | | | | 3.25 | 4 | 13 | 42.25 |
| 3.5 | 2 | 7 | 24.5 | 3.50 | 10 | 35 | 122.50 |
| | | | | 3.75 | 16 | 60 | 225.00 |
| 4.0 | 3 | 12 | 48.0 | 4.00 | 19 | 76 | 304.00 |
| | | | | 4.25 | 16 | 68 | 289.00 |
| 4.5 | 2 | 9 | 40.5 | 4.50 | 10 | 45 | 202.50 |
| | | | | 4.75 | 4 | 19 | 90.25 |
| 5.0 | 1 | 5 | 25.0 | 5.00 | 1 | 5 | 25.00 |
| $\Sigma$ | 9 | 36 | 147.0 | $\Sigma$ | 81 | 324 | 1309.50 |

Again $\mu_{\bar{x}} = \frac{324}{81} = 4$ and

$$\sigma^2_{\bar{x}} = \frac{1}{81} \times \left(1309.50 - \frac{324^2}{81}\right) = 0.1667 = \frac{\sigma^2}{n}$$

## The distribution of statistics

Consider two independent normal population, $N_1 = 2, n_1 = 3$, then $N_1^{n_1} = 8$; $N_2 = 3, n_2 = 2$, then $N_2^{n_2} = 9$. There are totally 72 differences $\bar{x}_1 - \bar{x}_2$

| $\bar{x}_1 - \bar{x}_2$ | times | $f(\bar{x}_1 - \bar{x}_2)$ | $f(\bar{x}_1 - \bar{x}_2)^2$ |
|---|---|---|---|
| 4 | 1 | 4 | 16 |
| 3 | 5 | 15 | 45 |
| 2 | 12 | 24 | 48 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| -1 | 12 | -12 | 12 |
| -2 | 5 | -10 | 20 |
| -3 | 1 | -3 | 9 |
| $\Sigma$ | 72 | 36 | 168 |

The mean $\mu_{\bar{x}_1-\bar{x}_2}$ and variance $\sigma^2_{\bar{x}_1-\bar{x}_2}$ of the distribution of the sample mean difference

$$\begin{aligned}
\mu_{\bar{x}_1-\bar{x}_2} &= \frac{\sum f(\bar{x}_1-\bar{x}_2)}{N_1^{n_1}N_2^{n_2}} = \frac{36}{72} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = \mu_1 - \mu_2 \\
\sigma^2_{\bar{x}_1-\bar{x}_2} &= \frac{1}{N_1^{n_1}N_2^{n_2}}\left\{\sum f(\bar{x}_1-\bar{x}_2)^2 - \frac{[f(\bar{x}_1-\bar{x}_2)]^2}{N_1^{n_1}N_2^{n_2}}\right\} \\
&= \sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}
\end{aligned}$$

The distribution of the sample mean difference from two normal population is also normal $\boxed{\mathcal{N}(\mu_1-\mu_2, \sigma^2_{\bar{x}_1-\bar{x}_2})}$

**Notice**: to estimate $\sigma^2$, the sample $n$ should be large enough. In case $\sigma^2$ unknown and $n < 30$, to use sample variance to estimate $\sigma^2$, $\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$ is no longer normal, but $t$-distribution with degree of freedom $\nu = n - 1$, here $\boxed{s_{\bar{x}} = \frac{s}{\sqrt{n}}}$ is the standard deviation of the sample mean:

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu}{2}}$$

Heavy tail distribution, with $\nu \to \infty$, $t \to \mathcal{N}$.

$$\mu_t = 0, \qquad \sigma_t^2 = \frac{\nu}{\nu - 2}$$

## $\chi^2$ distribution

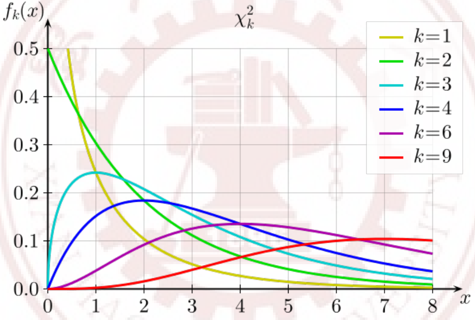
Suppose $u \sim \mathcal{N}(0,1)$, take $k$ i.i.d. samples $u_1, u_2, \cdots, u_k$, define

$$u = u_1^2 + \cdots + u_k^2 = \sum_{i=1}^{k} u_i^2$$

then $u \sim \chi_k^2$ where

$$p(u) = \frac{u^{k/2-1}}{2^{k/2}\Gamma(k/2)} \exp\left\{-\frac{u}{2}\right\}$$

where $x > 0$ if $k = 1$; otherwise $x \geq 0$




Mean: $k$; Variance: $2k$; Median: $\approx k\left(1-\frac{2}{9k}\right)^3$, Mode: $\max(k-2,0)$, $k \geq 30$, $\chi^2$ approximates normal.

## $F$ distribution

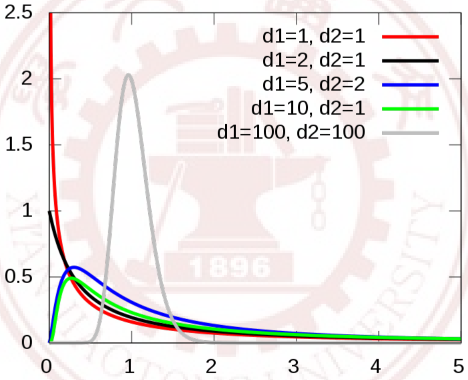
Two independent samples with sizes $n_1$ and $n_2$ from $\mathcal{N}(\mu, \sigma^2)$, their sample variances are $s_1^2$ and $s_2^2$, define

$$u = \frac{s_1^2}{s_2^2}$$

then $u \sim F$where

$$\begin{aligned} p(u; n_1, n_2) &= \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} n_1^{\frac{n_1}{2}} n_2^{\frac{n_2}{2}} \frac{u^{\frac{n_1}{2}-1}}{(n_1 u + n_2)^{\frac{n_1+n_2}{2}}} \\ &= \frac{1}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} u^{\frac{n_1}{2}-1} \left(1 + \frac{n_1}{n_2} u\right)^{-\frac{n_1+n_2}{2}} \end{aligned}$$

# The distribution of statistics




Mean: $\frac{n_2}{n_2-2}$ for $n_2 > 2$; variance $\frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}$ for $n_2 > 4$; Mode: $\frac{n_1-2}{n_1}\frac{n_2}{n_2+2}$ for $n_1 > 2$

**Definition**

Statistical inference is the process of using data analysis to deduce properties of an underlying probability distribution. Inferential statistical analysis infers properties of a population, for example by testing hypotheses and deriving estimates. It is assumed that the observed data set is sampled from a larger population.

Inferential statistics can be contrasted with descriptive statistics. Descriptive statistics is solely concerned with properties of the observed data, and it does not rest on the assumption that the data come from a larger population.

## Hypothesis testing

A statistical hypothesis, sometimes called confirmatory data analysis, is a hypothesis that is testable on the basis of observing a process that is modeled via a set of random variables.

A statistical hypothesis test is a method of statistical inference.

Commonly, two statistical data sets are compared, or a data set obtained by sampling is compared against a synthetic data set from an idealized model.

Hypothesis testing steps:

- A hypothesis is proposed for the statistical relationship between the two data sets, and this is compared as an alternative to an idealized null hypothesis that proposes no relationship between two data sets.
- The comparison is deemed statistically significant if the relationship between the data sets would be an unlikely realization of the null hypothesis according to a threshold probability — the significance level.
- Hypothesis tests are used in determining what outcomes of a study would lead to a rejection of the null hypothesis for a pre-specified level of significance.
- The process of distinguishing between the null hypothesis and the alternative hypothesis is aided by identifying two conceptual types of errors (type 1 & type 2), and by specifying parametric limits on e.g. how much type 1 error will be permitted.

An alternative framework for statistical hypothesis testing is to specify a set of statistical models, one for each candidate hypothesis, and then use model selection techniques to choose the most appropriate model.

The most common selection techniques are based on either Akaike information criterion or Bayes factor.

## Step I: Null and Alternative Hypothesis

Questions?