Concepts of Statistical Inference

Review: $(X_1,...,X_n)$ are *iid* distributed with $E(X_i) = \mu$, $Var(X_i) = \sigma^2$

Probability properties of $\overline{X}$ (mean and variance)

   a. $E(\overline{X}) = E(\frac{X_1 + ... + X_n}{n}) = \frac{1}{n}[E(X_1) + ... + E(X_n)] = \frac{n \cdot \mu}{n} = \mu$

   b. $Var(\overline{X}) = \frac{\sigma^2}{n}$.

Distributional properties of $\overline{X}$

   a. Exact distribution for a normal population

     when $X_i \sim^{iid} N(\mu, \sigma^2)$, then **for all** n

$$\overline{X} \sim^{exactly} N(\mu, \frac{\sigma^2}{n}) \quad \Leftrightarrow \quad \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim^{exactly} N(0,1)$$

   b. *Central limit theorem for **any** population* (中央極限定理)

$$\text{As } n \to \infty, \ \overline{X} \sim^{approximately} N(\mu, \sigma^2/n) \quad \Leftrightarrow \quad \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim^{approximately} N(0,1)$$

Normal approximation to Binomial （中央極限定理的應用）

$$Y = X_1 + ... + X_n \sim Binomial(n, p) \quad (X_i \sim^{iid} Bernoulli(p))$$

$$E(Y) = np \ \& \ Var(Y) = np(1-p)$$

$$\frac{Y - np}{\sqrt{np(1-p)}} = \frac{\overline{X} - p}{\sqrt{p(1-p)/n}} \sim^{approximately} N(0,1)$$

Continuation correction (不考)

$$\Pr(a \le Y \le b) = \sum_{y=a}^{y=b} \binom{n}{y} p^y (1-p)^{n-y}$$

$$\approx \Pr(a - 0.5 < Y < b + 0.5)$$

$$= \Pr(\frac{a - 0.5 - np}{\sqrt{np(1-p)}} < \frac{Y - np}{\sqrt{np(1-p)}} < \frac{b + 0.5 - np}{\sqrt{np(1-p)}})$$

$$\approx \Pr(\frac{a - 0.5 - np}{\sqrt{np(1-p)}} < Z < \frac{b + 0.5 - np}{\sqrt{np(1-p)}})$$

Ex: Assume that the probability that a NYCU student never plays online games is $p = 0.2$. Randomly select 100 NYCU students, find the probability that <u>at least 30</u> students have never played online games.

$Y \sim Bonomial(n = 100, p = 0.2)$

$$\Pr(Y \geq 30) = \sum_{x=30}^{100} \binom{100}{x}(0.2)^x(0.8)^{100-x}$$

$$= \Pr(\frac{Y - 100 \times 0.2}{\sqrt{100 \times 0.2 \times 0.8}} \geq \frac{30 - 20}{\sqrt{16}})$$

$$\approx \Pr(Z \geq 2.5) = 0.0062$$

Chebyshev's inequality. This inequality points out another useful property of the standard deviation. In particular, it states that "The probability that any random variable X falls within k standard deviations of its mean is at least $1 - 1/k^2$." For example, if we know that X has mean 3 and standard deviation 1, then we can conclude that the probability that X lies between 1 and 5 ($k = 2$ standard deviations from the mean) is at least $1 - 1/2^2 = .75$.

Chebyshev's Inequality

Let X be a random variable with $E(X) = \mu$ and $Var(X) = \sigma^2$.

For $k > 0$

$$\Pr(|X - \mu| \geq k\sigma) = \Pr(|X - \mu|^2 \geq k^2\sigma^2)$$

$$\leq \frac{E(|X - \mu|^2)}{k^2\sigma^2} \qquad .$$

$$= \frac{1}{k^2}$$

$$\Pr(|X - \mu| < k\sigma) = 1 - \Pr(|X - \mu| \geq k\sigma)$$

$$\geq 1 - \frac{E(|X - \mu|^2)}{k^2\sigma^2}$$

$$= 1 - \frac{1}{k^2}$$

(a) Let X denote the amount of rainfall received per week in a region. Assume that $\mu = 1.00$ inch and $\sigma = .25$ inch. Would it be unusual for this region to receive more than 2 inches of rain in a given week? Explain on the basis of Chebyshev's inequality.

X = amount of rainfall per week with $\mu = 1, \sigma^2 = \dfrac{1}{16}$

$\Pr(X > 2) = \Pr(X - \mu > 2 - 1)$

We know that by Chebyshev's inequality

$$\Pr(|X - \mu| > 1) = \Pr(|X - \mu|^2 > 1^2)$$
$$\leq \frac{Var(X)}{1}$$
$$= (\frac{1}{4})^2 = \frac{1}{16}$$

Also,

$$\Pr(|X - \mu| > 1)$$
$$= \Pr(X - 1 > 1) + \Pr(X - 1 < -1)$$
$$\leq \frac{1}{16} + 0 = \frac{1}{16}$$

$$\Pr(X > 2) \leq \frac{1}{16} - \Pr(X < 0)$$
$$\leq \frac{1}{16} = 0.0625$$

(b) Let X denote the number of cases of rabies reported in a given state per week. Assume that $\mu = \frac{1}{2}$ and $\sigma^2 = \frac{1}{25}$. Would it be unusual to observe two cases in a given week? Explain on the basis of Chebyshev's inequality.

b. X = number of rabies with $\mu = \dfrac{1}{2}, \sigma^2 = \dfrac{1}{25}$

$$\Pr(|X - \mu| > \frac{3}{2}) = \Pr(X - \frac{1}{2} > \frac{3}{2}) + \Pr(X - \frac{1}{2} < -\frac{3}{2})$$
$$= \Pr(X > 2) + \Pr(X < -1)$$
$$= \Pr(X > 2)$$

$$\Pr(|X - \mu| > \frac{3}{2}) = \Pr(|X - \mu|^2 > \frac{3^2}{2^2})$$
$$\leq \frac{Var(X)}{9/4}$$
$$= \frac{1}{25} \times \frac{4}{9} = \frac{4}{225} \approx 0.0178$$

3

Statistical Inference for μ

1. point estimation (點估計) (will not be on the exam)

   - $\hat{\theta}$: a statistic which is function of $(X_1,...,X_n)$ $\rightarrow$ a random variable

   - If $\hat{\theta}$ is used to estimate an unknown parameter θ, we say that $\hat{\theta}$ is a point

   estimator of θ.

   Ex: $\bar{X}$ is a point estimator of μ

Two important issues for point estimation: (will not be on the exam)

   - How to <u>construct</u> an estimator for estimating θ (i.e. μ)?

      * Maximum likelihood estimation (最大概似法)

      * Method of Moment (動差法)

   - How to <u>evaluate</u> the performance of $\hat{\theta}$ (i.e. $\bar{X}$)

      * Bias (偏誤): $E(\hat{\theta}) - \theta$ $\rightarrow$ validity

   $E(\bar{X}) - \mu = \mu - \mu = 0$ ($\bar{X}$ is an unbiased estimator of μ, 不偏估計量)
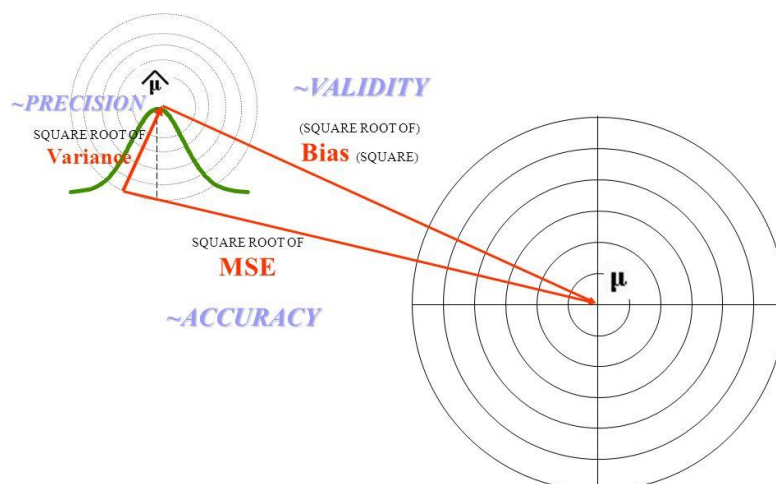
      * $Var(\hat{\theta})$ $\rightarrow$ precision

      * Mean squared errors (MSE):

   $E[(\hat{\theta} - \theta)^2] = [E(\hat{\theta}) - \theta]^2 + Var(\hat{\theta}) = \text{bias}^2 + Var(\hat{\theta})$

2. Interval estimation (區間估計): use an interval to estimate μ

3. hypothesis testing (假設檢定): test whether the value of μ equal to a hypothesized value μ_0.

* Construction of a point estimator (will not be on the exam)

Approach 1: Method of moment (動差法) → *proposed by Karl Pearson*

Principle: If θ can be expressed in terms of moments of X_i,

then use sample moments to estimate the population moments

Sample moments ($\hat{\theta}$)	Population moments (θ)
$\sum_{i=1}^{n} X_i / n$	$E(X_i)$
$\sum_{i=1}^{n} X_i^k / n$	$E(X_i^k)$

Approach 2: Maximum likelihood estimation (最大概似法)→ *proposed by RA Fisher*

From the aspect of probability, the value of θ is given

$$\Pr(X_1 = x_1, ..., X_n = x_n)$$
$$= \prod_{i=1}^{n} \Pr(X_i = x_i \mid \theta)$$

= the joint probability of obtaining the observed sample $(x_1, ..., x_n)$
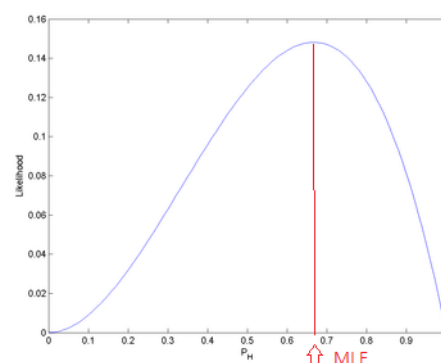
→ $L(\theta) = \prod_{i=1}^{n} f_\theta(x_i)$

= the likelihood function of $(x_1, ..., x_n)$ with θ being unknown

→ reveals the evidence of the unknown parameter value provided by observed data

The *maximum likelihood estimator (MLE)* is obtained by maximizing

$$L(\theta) = \prod_{i=1}^{n} f_\theta(x_i)$$

or $\qquad l(\theta) = \sum_{i=1}^{n} \log\{f_\theta(x_i)\}$

Q: How to obtain the formula of MLE? A: Solve the following equation:

$$\frac{\partial l(\theta)}{\partial \theta} = \frac{\partial \sum_{i=1}^{n} l_{\theta}(x_i)}{\partial \theta} = 0$$
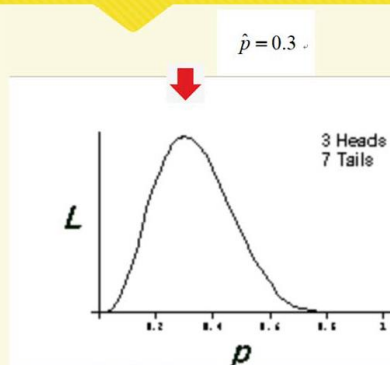
參數估計: 擲銅板得到正面的機率 (p)

Bernoulli trial:
$$X_i = 1 \text{ if head}; \quad X_i = 0 \text{ if tail}.$$
Data: $(x_1, ..., x_n)$
$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \quad \rightarrow \text{likelihood function}$$
$\hat{p}$ maximize $L(p)$

Maximum likelihood estimation

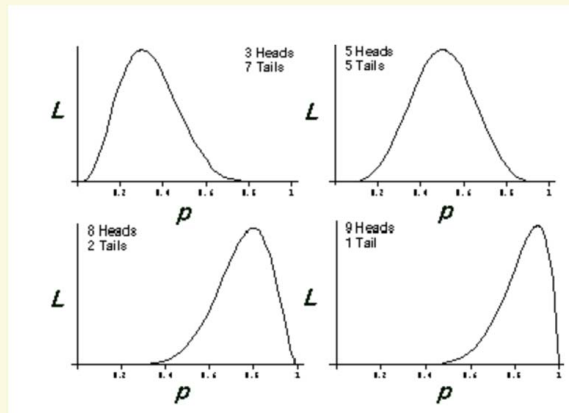$\hat{p} = 0.3$

3 Heads
7 Tails

$$\text{Max} \cdot L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

Maximize
$$\log L(p) = \sum_{i=1}^{n} \{x_i \log p + (1-x_i)\log(1-p)\}$$

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \# \text{of heads} / \# \text{of tosses}$$

隨機實驗每次結果不同!

Thus $\hat{p} = \sum_{i=1}^{n} X_i / n$ is a random variable.
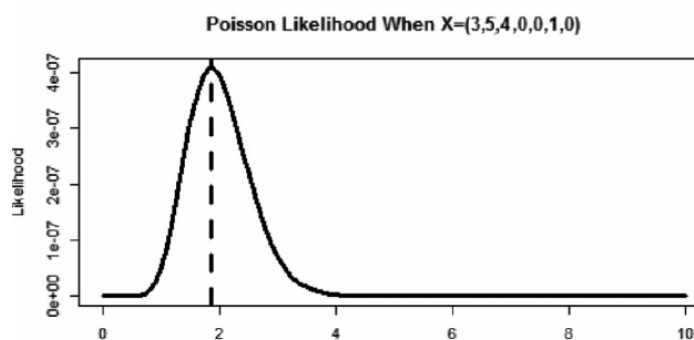
Example 1: random sample from a Poisson distribution

$$f_\mu(x) = \Pr(X = x) = \frac{e^{-\mu}\mu^x}{x!} \quad E(X_i) = \mu; \ Var(X_i) = \mu,$$

Likelihood function of μ

$$L(\mu) = \Pr(X_1 = x_1, ..., X_n = x_n) = \prod_{i=1}^{n} \Pr(X_i = x_i) = \prod_{i=1}^{n} \frac{e^{-\mu}\mu^{x_i}}{x_i!} = e^{-n\mu}\mu^{\sum_{i=1}^{n} x_i}\left(\prod_{i=1}^{n} x_i!\right)^{-1}$$

$$l(\mu) = -n\mu + (\sum_{i=1}^{n} x_i)\log(\mu) - \log\left(\prod_{i=1}^{n} x_i!\right)$$

$$\frac{\partial l(\mu)}{\partial \mu} = -n + \frac{\sum_{i=1}^{n} x_i}{\mu} = 0 \ \rightarrow \text{the MLE solution} \ \hat{\mu} = \bar{X}$$

Poisson Likelihood When X=(3,5,4,0,0,1,0)

* Method of Moments

First moment: $E(X_i) = \mu$

Use $\hat{\mu} = \bar{X}$ → $E(X_i) = \mu$

Use variance to estimate $\hat{\mu} = \sum_{i=1}^{n}(X_i - \bar{X})^2 / n$ → $Var(X_i) = \mu$

Remark: The method of moments will not produce a unique result.

* Maximum likelihood estimation

Example: random sample from a normal distribution

$X_i \sim^{iid} N(\mu, \sigma^2)$ → $E(X_i) = \mu$; $Var(X_i) = \sigma^2$

Use the first moment to estimate μ → $\hat{\mu} = \bar{X}$

Use the first two moments to estimate σ^2 → $\hat{\sigma}^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 / n$

Likelihood function of (μ, σ^2)

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-(\frac{X_i - \mu}{\sigma})^2] = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp[\frac{-\sum_{i=1}^{n}(X_i - \mu)^2}{2\sigma^2}].$$

$$l(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{2\sigma^2}$$

$$\frac{\partial l(\mu, \sigma^2)}{\partial \mu} = 0 \ \& \ \frac{\partial l(\mu, \sigma^2)}{\partial \sigma^2} = 0 \leftrightarrow \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{2\sigma^4} = \frac{n}{2}\frac{1}{\sigma^2}$$

Solution: $\hat{\mu} = \bar{X}$;

$$\hat{\sigma}^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 / n$$

Evaluating the performance of a point estimator (will not be on the exam)

Let $\hat{\theta}$ be an estimator of θ.

Note: $\hat{\theta}$ is a random variable since it is a function of the random sample $X_1,...,X_n$

Criteria

There are some criteria for evaluating the performance of $\hat{\theta}$.

1. Bias of $\hat{\theta}$ = $E(\hat{\theta}) - \theta$ （偏誤）

   If $E(\hat{\theta}) = \theta$, then $E(\hat{\theta}) = \theta$ is an unbiased estimator of $E(\hat{\theta}) = \theta$.
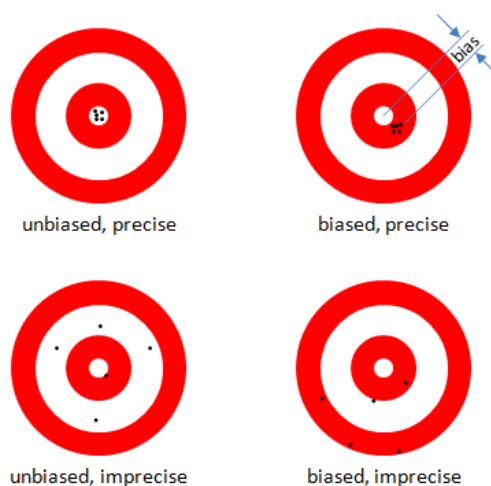
Examples

$E(\bar{X}) = \mu$ → unbiased

$E[\sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)] = E(S^2) = \sigma^2$ → S^2 is an unbiased estimator of σ^2

$E[\sum_{i=1}^{n}(X_i - \bar{X})^2/n] = E(\tilde{S}^2) \neq \sigma^2$ → $\tilde{S}^2$ is NOT an unbiased estimator of σ^2

2. Variance of $\hat{\theta}$ → evaluate the precision of $\hat{\theta}$

   Small $Var(\hat{\theta})$ implies that the values of $\hat{\theta}$ obtained from different sampling

   results are close to each other.

unbiased, precise   biased, precise

unbiased, imprecise   biased, imprecise

3. Mean squared error (MSE)

$$E[(\hat{\theta} - \theta)^2] = Var(\hat{\theta}) + \left[E(\hat{\theta} - \theta) \right]^2$$

$$= \text{variance of } \hat{\theta} + \text{the expected value of squared bias}$$

Comparing two estimators based on <u>relative efficiency</u>

$$\frac{E[(\hat{\theta}_1 - \theta)^2]}{E[(\hat{\theta}_2 - \theta)^2]} = \frac{MSE(\hat{\theta}_1)}{MSE(\hat{\theta}_2)} \qquad \text{(MSE: the smaller, the better)}$$

Evaluation of $\bar{X} = \sum_{i=1}^{n} X_i / n$

1. $E(\bar{X}) = \mu$ → unbiased

2. $Var(\bar{X}) = \dfrac{\sigma^2}{n}$ → attains the lower bound (the best choice)

3. $MSE(\bar{X}) = \dfrac{\sigma^2}{n} + 0$

Remarks:

- $\bar{X} = \sum_{i=1}^{n} X_i / n$ is the best estimator of μ

- However in many other applications, there may not exist the best choice.

Topic: Confidence Interval for μ

Point estimator of μ : $\bar{X}$

Interval estimator of μ : $\bar{X} \pm$ margin of error

Objective: use a <u>random interval</u> $(L(X_1,...,X_n), U(X_1,...,X_n))$ to estimate μ

Rule: fix the confidence level such that

$$\Pr(\mu \in L(X_1,...,X_n), U(X_1,...,X_n)) = 1 - \alpha$$

Confidence level (信心水準) = $(1-\alpha)100\%$

Remarks:

- $U(X_1,...,X_n) = \bar{X} +$ margin of error $\rightarrow$ the upper bound

- $L(X_1,...,X_n) = \bar{X} -$ margin of error $\rightarrow$ the lower bound

- The formula for the margin of error (抽樣誤差範圍) depends on

    a. the sampling distribution of $\bar{X}$

    b. whether $Var(X) = \sigma^2$ is known.

- *Convention:* 95% confidence level $\Leftrightarrow$ $1 - \alpha = 0.95$ $\Leftrightarrow$ $\alpha = 0.05$

Condition 1: $X_i \sim^{iid} N(\mu, \sigma^2)$ with $Var(X) = \sigma^2$ being known

Formula of $(1-\alpha)\cdot 100\%$ **confidence interval for** μ **with** σ^2 **known**

$$\bar{X} \pm \text{ margin of error} \Leftrightarrow \bar{X} \pm z_{\alpha/2} \cdot \sigma / \sqrt{n}$$
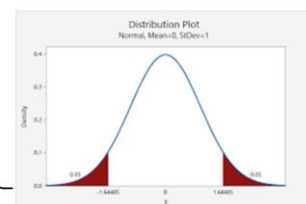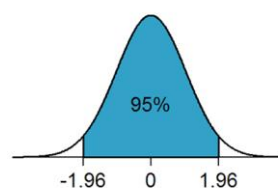
$$\Leftrightarrow [\bar{X} - z_{\alpha/2} \cdot \sigma / \sqrt{n}, \bar{X} + z_{\alpha/2} \cdot \sigma / \sqrt{n}],$$

where $z_{\alpha/2}$ satisfies $\Pr(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1 - \alpha$ or $\Pr(Z > z_{\alpha/2}) = \alpha/2$

Useful table values:

95% CI ($\alpha = 5\%$): $z_{0.025} = 1.96$;

90% CI ($\alpha = 10\%$): $z_{0.05} = 1.645$

Derivations of the formula

The first step: Distribution theory

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

Then $$\Pr(-z_{\alpha/2} \le \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \le z_{\alpha/2}) = 1 - \alpha$$

The second step: rearrange the inequality

$$\Pr(\mu \in [\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}]) = 1 - \alpha$$

Note: You need to be familiar with the following techniques

- $|x - a| \le k \Leftrightarrow -k \le x - a \le k$

- $|x - a| > k \Leftrightarrow x - a > k$ or $x - a < -k$

Remarks about $(1 - \alpha)100\%$ **confidence interval for** μ :

$$[\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}]$$

a. Before sampling, $[\bar{X} - z_{\alpha/2} \cdot \sigma / \sqrt{n}, \bar{X} + z_{\alpha/2} \cdot \sigma / \sqrt{n}]$ is a random interval and

$1 - \alpha$ represents that the probability that μ is contained in the interval.

b. In reality, the data that we obtain is the result (realization) after the sampling.

Hence $[\bar{X}_{obs} - z_{\alpha/2} \cdot \sigma / \sqrt{n}, \bar{X}_{obs} + z_{\alpha/2} \cdot \sigma / \sqrt{n}]$ is a fixed interval and not random any

more.

c. Accordingly, $1 - \alpha$ is called as the "confidence level" rather than "probability".

Margin of error (誤差範圍) = $\pm z_{\alpha/2} \cdot \sigma / \sqrt{n}$ → the smaller, the better

The margin of errors are affected by

        a. σ: population variability (small σ, 天生麗質)

        b. n: sample size (large n, 後天努力)

        c. $z_{\alpha/2}$: determined by the confidence level (the set standard).

        (Lower the standard such as 95% → 90%, the interval will be narrower).

Given the margin of error, how to find the smallest sample size

Solve the inequality with n being unknown

$$z_{\alpha/2} \cdot \sigma / \sqrt{n} \leq K$$

The margin of error in surveys: $\pm$ 3%

$$\hat{p} = \sum_{i=1}^{n} B_i / n \quad (B_i = 1 \Leftrightarrow \text{ the ith person votes for the candidate})$$

$$\sqrt{Var(\hat{p})} = \sqrt{\frac{p(1-p)}{n}} \leq \frac{0.5}{\sqrt{n}} \quad \text{(the largest variance happens when } p = 0.5)$$

For 95% confidence level, take

$$z_{\alpha/2} = z_{0.025} = 1.96$$

Margin of errors = $\pm$ 0.03

→ $1.96 \times \sqrt{Var(\hat{p})} \leq 0.03$

→ $1.96 \times \dfrac{0.5}{\sqrt{n}} \leq 0.03$

→ $\left(1.96 \times 0.5 / 0.03\right)^2 \leq n$

→ $32.5 \leq \sqrt{n}$

→ $1057 \leq n$

Story: Rivalry between Pearson and Fisher

統計人物

method of moment
correlation coefficient
Chi-squared test

1857 Karl Pearson 1936

T Distribution

1876 William Sealy Gosset 1937

maximum likelihood estimation
exprimental design/ ANOVA
Time Series
Discriminant analysis
the greatest of Darwin;s successors

1890 Ronald Aylmer Fisher 1962

PROFESSOR KARL PEARSON AND THE METHOD OF MOMENTS

By R. A. FISHER, Sc.D., F.R.S.

I. APOLOGY

SHORTLY before his death the late Prof. Karl Pearson wrote a paper on the method of moments, which was published in his journal, *Biometrika*, for June 1936. The paper is unfortunately marred by great bitterness, and by vehement attacks on an Indian writer, R. S. Koshal, whose offence appears to be that of deciding, after trial, that a curve of Pearson's type I could be fitted to his data more successfully by the method of maximum likelihood than by the method of moments. Pearson's allusions to Koshal's work, some of which will have to be quoted, are so unjust as to be quite inexplicable. His paper, however, deserves attention as an authentic exposition of the manner in which followers of Prof. Pearson should proceed in fitting frequency curves, and as the first attempt to answer the doubts and difficulties which have been felt by others with respect to the methods he had previously put forward. I hope later to publish Dr Koshal's further examination of the problem, and for the present shall confine attention to Prof. Pearson's methods.

Though the occasion of this paper is Pearson's attack on Koshal, it has been impossible to treat the matter in due perspective without a general criticism of methods originating with Pearson, which have been widely disseminated. The intrinsic worth of those methods has long appeared to me to have been gravely exaggerated. Pearson opens his paper with the italicized query "*Wasting your time fitting curves by moments, eh?*" thus expressing in his own words and style the scepticism with which he felt his procedures were being regarded by others. The question he raised seems to me not at all premature, but rather overdue. During his last years Pearson's intimates have earnestly represented that irritation and controversy might be dangerous to his health. In consequence, the discussion of many points has been suspended. With Pearson's death my obligation to examine frankly the status of the Pearsonian methods reasserts itself, and in the last section I have attempted a brief review of the situation as it now stands.

Moment-based estimation methods are still popular in biomedical applications

GEEs belong to a class of regression techniques that are referred to as semiparametric because they rely on specification of only the first two moments. They are a popular alternative to the likelihood–based generalized linear mixed model which is more sensitive to variance structure specification.

國家衛生研究院第六任院長由梁賡義院士擔任

新任院長梁賡義院士為享譽國際的生物統計學者，在生物統計、公共衛生領域貢獻多年，學術成就與聲望廣獲國際肯定。梁院士畢業於國立清華大學理學院數學系，隨後於美國華盛頓大學公衛學院生物統計所取得博士學位。其任職約翰霍普金斯大學教授多年，研究領域為生物統計學以及流行病學。梁院士於1986年在約翰霍普金斯大學時與同事Scott Zeger設計了「廣義估計公式」，廣泛運用在臨床及公共衛生研究領域。1995年獲選為美國統計學會會員，2002年當選中央研究院院士，2012年當選世界科學院院士，2015年當選美國國家醫學院院士。

https://udn.com/news/story/7266/5016759

台灣之光！國衛院院長梁賡義獲選全球50大改變創造者

2020-11-15 14:20 聯合報 / 記者邱宜君／台北即時報導

梁賡義在美國約翰霍普金斯大學任教期間與其同事，在 1986 年共同發表「廣義線性模式於縱向資料分析」，突破了過去的統計限制，不需花漫長時間等待病情惡化或死亡，只要分析數據在一段時間內的變化，就可推測未來的風險。這項創新方法過去 34 年來已被發揚光大，廣泛應用於公衛、醫藥、遺傳、社會科學等。

梁賡義表示，發明這個方法的契機，是當時有位教授正在研究，母親壓力與孩子身體健康之間的關係，需要每天記錄母子各自的身心狀況。由於孩子的成長從懷孕期間就與母親有關，同一家戶內的孩子與孩子之間也有相關，孩子又會各自隨著時間成長。要隨著時間軸進行縱向觀察，又要同時處理資料之間不完全獨立的問題，當時的可用的統計分析方法面臨諸多限制。

因此華盛頓大學畢業的梁賡義與另一位從普林斯頓畢業的同事，一起研發出了這個適合分析不互相獨立之縱向資料的方法。梁賡義表示，以前研究新藥有沒有效，要看兩組不同的人，分別吃新藥劑和舊藥，並且長時間觀察他的疾病惡化及死亡率。有了這方法之後，只要能夠找到關鍵的生物指標，就不需要等待很長時間，在比較短時間內重複測定指標的變化，就能夠推測未來病情的走向。

2018 高雄市長選舉
《三立民調》高雄市長選情激烈 陳其邁領先韓國瑜 1.1%

根據《三立新聞》最新出爐的高雄民調，年底高雄市長支持度，陳其邁以 44.3% 領先，韓國瑜 支持度 43.2% 緊追在後，至於另外兩名無黨籍的蘇盈貴及璩美鳳的支持度分別為 0.8%、0.7%，沒有無明確意見 11.1%。陳其邁僅領先韓國瑜 1.1%，呈現五五波態勢。

此次調查對象為戶籍在高雄市且年滿 20 歲具有投票權之民眾，**調查方法係採用電話調查隨機抽樣方式進行**，調查時間為 107 年 11 月 3 日至 11 月 4 日。抽樣方法採分層比率隨機抽樣方式進行。**共完成 1,023 份有效樣本，在 95%信心水準下，抽樣誤差為正負 3.1 個百分點。**

聯合報高雄市長民調／韓國瑜 49%領先陳其邁 32% 2018-11-12

本報高雄市長選情調查發現，「韓流」翻轉高雄選情，韓國瑜不論就支持度或看好度皆超前，韓支持度攀升至四成九，領先陳其邁的三成二；看好韓有機會入主市府的選民上升到四成九，覺得陳有勝算者則降到二成六。這次調查於十一月八日至十日晚間進行，成功訪問了一千零七十四位設籍高雄市的成年民眾，另有一百七十四人拒訪；**在百分之九十五的信心水準下，抽樣誤差在正負三點零個百分點以內**。調查以高雄市住宅電話為母體作尾數兩位隨機抽樣，依據高雄市成年民眾之性別、年齡及區域結構等進行加權。調查經費來源和計畫主持人皆為聯合報社。

2018年高雄市市長選舉結果					[合併]
號次	候選人	政黨	得票數	得票率	當選標記
1	韓國瑜	中國國民黨	892,545票	53.87%	⊘
2	陳其邁	民主進步黨	742,239票	44.80%	
3	璩美鳳	無黨籍	7,998票	0.48%	
4	蘇盈貴	無黨籍	14,125票	0.85%	
選舉日期	2018年11月24日		選舉人數	2,281,338人	
投票率	73.54%		投票人數	有效：1,656,907人 無效：20,743人	

2014-12-01 民調失準檢討

〔即時新聞／綜合報導〕九合一大選結果出爐後，讓藍綠雙方跟選民都感到相當訝異，因為許多縣市的選舉結果都跟先前的民調落差很大，許多人說沒想到國民黨這次會輸這麼多，民調公司則表示，因這次選舉出現大量的青年票，較難預測才導致民調失準。

據 TVBS 新聞報導，選前民調公司預估台北市柯文哲會贏對手 13 到 19 個百分點，選舉結果則是柯文哲勝 16％，還算準確，**而新北市原本預估朱立倫會大贏對手游錫堃近 20％，沒想到最後卻只小贏 1.28％，讓許多人跌破眼鏡，桃園市則看好吳志揚會贏鄭文燦 8 到 24％，結果卻是情況翻轉，反倒是鄭文燦贏了近 3%**

這次的九合一選舉，許多地區出現民調失準的現象，對此，民調公司表示，經過太陽花學運後，年輕人的投票意願及比例遠超過以往估計，反倒是以往投票比例較高的 50 歲以上族群，在這次投票比例中大幅減少，導致選票結構改變，加上網路的發展，讓一貫使用室內電話進行調查的民調，無法預測到年輕世代的想法，才會導致這次民調與選後結果的大幅落差。

2020 總統大選

委託調查單位	調查時間	有效樣本	民進 蔡英文 / 民進 賴清德	國民 韓國瑜 / 無黨 張善政	親民 宋楚瑜 / 無黨 余 湘
綠黨（頁面存檔備份，存於網際網路檔案館）	12-29 - 12-30	1,044	54.2%	20.8%	6.0%
自由時報（頁面存檔備份，存於網際網路檔案館）	12-23 - 12-25	1,074	54.25%	15.59%	4.76%
台灣民意基金會（頁面存檔備份，存於網際網路檔案館）	12-23 - 12-24	1,075	52.5%	21.9%	9.5%

← 2016年	2020年1月11日	2024年 →
投票率	74.90%（▲8.63%）	

	宋楚瑜	韓國瑜	蔡英文
獲提名人	宋楚瑜	韓國瑜	蔡英文
政黨	親民黨	中國國民黨	民主進步黨
競選搭檔	余湘（無黨籍）	張善政（無黨籍）	賴清德（民主進步黨）
民選得票	608,590	5,522,119	8,170,231
得票率	4.26%	38.61%	57.13%

2016 general election polls

Clinton vs. Trump - tracking polls

Clinton-Trump 2016 head-to-head tracking polls (November 2016)					[hide]
Poll	🔵 Hillary Clinton	🔴 Donald Trump	Unsure or Other	Margin of Error	Sample Size
IDB/TIPP tracking poll⌐ November 4-7, 2016	43%	42%	15%	+/-3.1	1,107

← 2012	**November 8, 2016**	2020 →

538 members of the Electoral College
270 electoral votes needed to win

Turnout 55.7%[1] ▲ 0.8 pp

	Donald Trump	Hillary Clinton
Nominee	Donald Trump	Hillary Clinton
Party	Republican	Democratic
Home state	New York	New York
Running mate	Mike Pence	Tim Kaine
Electoral vote	304[a]	227[a]
States carried	30 + ME-02	20 + DC
Popular vote	62,984,828	65,853,514
Percentage	46.1%	48.2%

Why 2016 election polls missed their mark

BY ANDREW MERCER, CLAUDIA DEANE AND KYLEY MCGEENEY

NOVEMBER 9, 2016

The results of Tuesday's presidential election came as a surprise to nearly everyone who had been following the national and state election polling, which consistently projected Hillary Clinton as defeating Donald Trump. **Relying largely on opinion polls, election forecasters put Clinton's chance of winning at anywhere from 70% to as high as 99%**, and pegged her as the heavy favorite to win a number of states such as Pennsylvania and Wisconsin that in the end were taken by Trump.

How could the polls have been so wrong about the state of the election?

One likely culprit is what pollsters refer to as nonresponse bias. This occurs when certain kinds of people systematically do not respond to surveys despite equal opportunity outreach to all parts of the electorate. We know that some groups – including the less educated voters who were a key demographic for Trump on Election Day – are consistently hard for pollsters to reach. It is possible that the frustration and anti-institutional feelings that drove the Trump campaign may also have aligned with an unwillingness to respond to polls. The result would be a strongly pro-Trump segment of the population that simply did not show up in the polls in proportion to their actual share of the population.

Some have also suggested that many of those who were polled simply were not honest about whom they intended to vote for. The idea of so-called "shy Trumpers" suggests that support for Trump was socially undesirable, and that his supporters were unwilling to admit their support to pollsters. This hypothesis is reminiscent of the supposed "Bradley effect," when Democrat Tom Bradley, the black mayor of Los Angeles, lost the 1982 California gubernatorial election to Republican George Deukmejian despite having been ahead in the polls, supposedly because voters were reluctant to tell interviewers that they were not going to vote for a black candidate.

A third possibility involves the way pollsters identify likely voters. Because we can't know in advance who is actually going to vote, pollsters develop models predicting who is going to vote and what the electorate will look like on Election Day. This is a notoriously difficult task, and small differences in assumptions can produce sizable differences in election predictions. We may find that the voters that pollsters were expecting, particularly in the Midwestern and Rust Belt states that so defied expectations, were not the ones that showed up. Because many traditional likely-voter models incorporate measures of enthusiasm into their calculus, 2016's distinctly unenthused electorate – at least on the Democratic side – may have also wreaked some havoc with this aspect of measurement.

When the polls failed to accurately predict the British general election in May 2015, it took a blue ribbon panel and more than six months of work before the public had the results of a data-driven, independent inquiry in hand. It may take a similar amount of time to get to the bottom of this election as well. The survey industry's leading standards association, the American Association for Public Opinion Research, already has an ad hoc committee in place to study the election and report back in May.

A supporter of Democratic U.S. presidential nominee Joe Biden holds signs while waiting outside Philadelphia convention center, a vote counting center for the 2020 U.S. presidential election, in Philadelphia, Pennsylvania, U.S., November 5, 2020. **(REUTERS)**

THE WALL STREET JOURNAL.

What went wrong with polls this year

5 min read . Updated: 06 Nov 2020, 05:52 AM IST
The Wall Street Journal

https://www.livemint.com/news/world/what-went-wrong-with-polls-this-year-11604621831841.html

Tuesday's election results delivered a second black eye for the nation's pollsters in as many presidential contests as well as the unmistakable message that they have misjudged their ability to measure political opinion in an era of cellphones, polarization and Donald Trump. (手機年代，意見極化，川普 → 民調困難)

With the vote-count still in progress, the size of this year's polling error is still unknown. **But as with the 2016 election, both national and state surveys left the impression that Election Night would bring clear Democratic gains, not a cliff-hanger.** (2016, 2020 都以為民主黨會大勝)

No matter the outcome in the Electoral College, Democratic presidential nominee Joe Biden is likely to hold his lead in the popular vote, which stood at 2 percentage points as of Wednesday afternoon. Many analysts expect that lead to grow as Democratic-leaning states in the West complete their tallies. Unknown, however, is whether it will grow large enough to match the lead Mr. Biden held

in the averaged, final results of national polls compiled by the websites RealClearPolitics or Fivethirtyeight.com, which on Election Day stood at 7.2 points and 8.4 points, respectively. **The final Wall Street Journal/NBC News poll reported a Biden lead of 10 points.** 民調認為拜登領先 10%

			■ Won	■ Leads	
Candidates		Electoral votes	Vote %	Vote count	
Joe Biden Democratic Party		290	50.6%	75,198,127	
Donald Trump Republican Party		214	47.7%	70,804,457	

While the final result may fall within the margin of error of national polls, opinion researchers said the outcome pointed to the need for evaluating their methods. (民調方法錯誤?)

"We're all trying to figure out where we go from here," said Mark Blumenthal, a veteran pollster who consults with opinion research firms.

Polls of many House and Senate races—and the election analysts who rely on them—also **appeared to be off target**, as predicted Democratic gains in both chambers of Congress failed to come to pass. 參眾兩院都估錯

Some pollsters had initial impressions about the cause of the missteps but said they needed to know more about the final results before digging in.

One theory gaining new attention was that renewed efforts this year to draw in more rural voters and those without four-year college degrees—groups supportive of Mr. Trump—were helpful but insufficient.

Some have come to believe that a distrust of institutions is more pervasive than anticipated across many voter groups, and that it leads conservative voters, even those with college degrees and urban addresses, to avoid participating in polls in disproportionate numbers. If so, the problem likely can't be corrected by adding more members of any one demographic group to a polling sample, they said. 對機構的不信任感導致部分選民不想參與民調

"I readily admit that there were problems this year, but it is too soon to know the extent of the problems, or what caused them,'' said Courtney Kennedy, who

supervises poll methodology for the Pew Research Center. In both 2016 and 2020, she said, "there was a widespread overstatement of Democratic support," but the causes this time around could be different. 兩次都高估民主黨選票,但原因不同

Some pollsters said that state and local polls had been particularly off-course. "District-level polling has rarely led us—or the parties and groups investing in House races—so astray," wrote analyst David Wasserman of the Cook Political Report, whose review of private polling by political groups had led him to predict Democratic gains in the House of between 10 and 15 seats, rather than the GOP gains that now seem likely.

But national polls also faced challenges. The final Journal/NBC News poll, taken in the past week of the campaign, found **Mr. Biden leading by 20 points among women and 23 points among seniors**—numbers that seemed to require Mr. Trump to put together an improbable coalition of voters to win.
NBC 錯得離譜: 以為拜登在女/年長者都有優勢, 事實上前者優勢沒那麼多, 後者(年長者)反而輸
By contrast, a large survey of the 2020 electorate called AP VoteCast found that **Mr. Biden won among women by 11 percentage points and that he lost seniors by 3 points**, rather than winning the group.

Moreover, the Journal/NBC poll appeared to understate the number of new voters that GOP recruiting efforts were bringing into the electorate, as well as their tilt toward Mr. Trump in some swing states, said Jeff Horwitt, a Democratic pollster who conducts the survey with Republican Bill McInturff.

In other words, Mr. Trump won several key states in the election in part because he changed the electorate, a possibility the pollsters had signaled but which the poll's results didn't fully reflect.

 "First-time voters are hard to reach for the same reason that you have to devote resources to get them to vote for the first time—they're harder to engage,'' said Mr. Horwitt. He added that he wanted to see final, nationwide results to better assess the performance of the poll. 首投族難以接觸到

Pollsters were grappling with whether the 2020 results showed new sources of

error, or whether problems known from 2016 hadn't been fixed.

A panel of the polling industry's professional group, the American Association for Public Opinion Research, said that one source of problems in 2016 was that an unusually large group of voters—some 13% in several swing states—only settled on a candidate in the final days. That meant pollsters missed those voters' late swing to Mr. Trump. 搖擺州有 13% 到最後幾天才決定

AP VoteCast, the large survey of the 2020 electorate, found only 5% of voters choosing a candidate in the final days, and they broke more modestly for Mr. Trump.

Another problem the panel identified: Some pollsters of swing states in 2016 were using methods that understated the voices of white, working-class voters and likely of rural voters. Some pollsters adopted statistical and other measures aimed at correcting the problem, but some didn't.

2016 年難以接觸到男性白人藍領階級，有的在 2020 有調整，有的沒有

Polling continues to face the challenge of declining participation. As fewer people pick up the phone for unknown callers, survey-response rates have fallen from 36% in 1997 to 6% in 2018 at the Pew Research Center, the group says. The rise of cellphones has raised costs, since by law they can't be called with autodialers, prompting pollsters to hunt for new ways to find respondents.

Mr. Blumenthal and others were skeptical that polling faced a significant problem this year from "shy Trump voters"—the idea that some voters, when interviewed, believe that it is socially unacceptable to say they support Mr. Trump.

The bulk of the problem, he said, is less likely to be that people are misleading pollsters than getting the right mix of people on the phone to begin with. That is likely to lead pollsters to think harder about who isn't participating in polls, and whether a meaningful number of people across many demographic groups, who in the main are conservative, decline to participate. While many pollsters monitor the share of Republicans and Democrats who participate and decline to participate, "it may be that the Republican voters you interview are on the more moderate side, whereas the ones you didn't interview might be more staunchly part of the Republican base,'' said Ms. Kennedy.

It is a hard problem to study, Mr. Blumenthal said. "**We don't interview the people who don't respond,'' he said. "We know much less about the voters we cannot reach than the ones we interview.''** 有些選民民調根本接觸不到