Lecture 8 11/4/2020

1. Review: from probability to statistics

Random sample:

$X_1,\ldots,X_n$ form a random sample (隨機樣本) from a population with mean μ and variance σ^2. We can write $X_i \overset{iid}{\sim} E(X_i)=\mu,\ Var(X_i)=\sigma^2$.

Properties of $\bar{X}$:

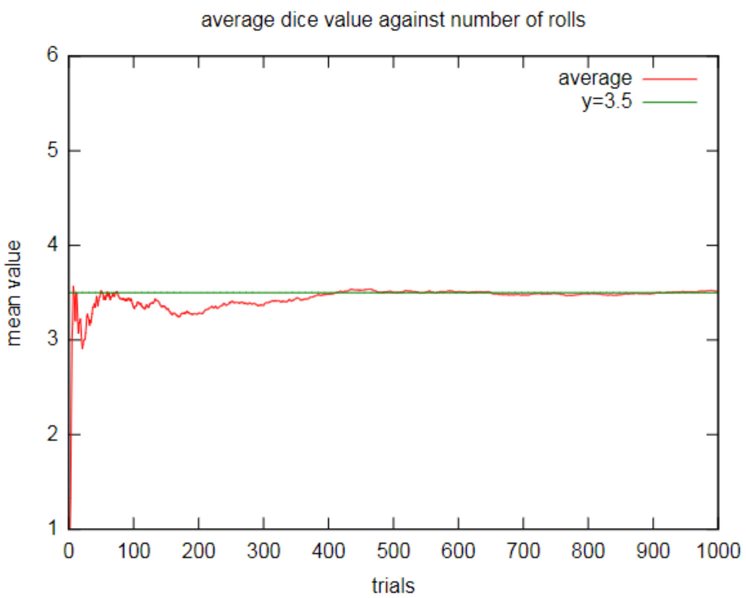
a. $E(\bar{X})=\mu$

b. $Var(\bar{X})=\dfrac{\sigma^2}{n}$, $\sqrt{Var(\bar{X})}=\sigma_{\bar{X}}=\dfrac{\sigma}{\sqrt{n}}$

Note: Taking the average of repeated measurements can reduce the error.

Law of Large Number (L.L.N., 大數法則,)

$$\bar{X}\to\mu \quad \text{as} \quad n\to\infty.$$



Remarks:

- You can see that when $n\to\infty$, $Var(\bar{X})=\dfrac{\sigma^2}{n}\to 0$.

- There are two versions of LLN: strong law and weak law.

- $\sum_{i=1}^{n} g(X_i)/n \to E[g(X)]=\int g(x)f(x)dx$

Central Limit Theorem: $X_i \sim^{iid} \quad E(X_i) = \mu, \; Var(X_i) = \sigma^2$.

$$\bar{X}_n \underset{n\to\infty}{\sim} N(\mu, Var(\bar{X}_n) = \sigma^2 / n) \Leftrightarrow \frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} \underset{n\to\infty}{\sim} N(0,1)$$

Graphical example:

There are 5 different types of random variables, their sample averages (n = 5) become very alike.

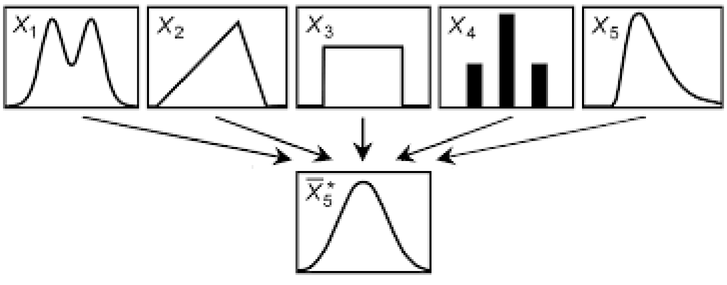


Exhibit 3.30: The central limit theorem is illustrated in the case of fiv arbitrarily selected independent random variables. Random variables X_1 X_2, X_3, and X_5 are continuous, so their PDFs are shown; X_4 is discrete so its PF is shown. The normalized average $\bar{X}_5^*$ is approximately $N(0,1$ All graphs indicate the interval [–3,3] on the *x*-axis.

Random Variable	Mean	Standard Deviation	Skewness	Kurtosis	Description
X_1	0.00	1.00	0.00	1.89	continuous
X_2	0.00	1.00	–0.41	2.41	continuous
X_3	0.00	1.00	0.00	1.80	continuous
X_4	0.00	1.00	0.00	2.00	discrete
X_5	0.00	1.00	1.62	7.89	continuous
$\bar{X}_5^*$	0.00	1.00	0.11	3.03	continuous

Comparison: LLN and CLT

Law of Large Number:

$$\bar{X}_n \underset{n\to\infty}{\to} \mu \Leftrightarrow \bar{X}_n - \mu \underset{n\to\infty}{\to} 0$$

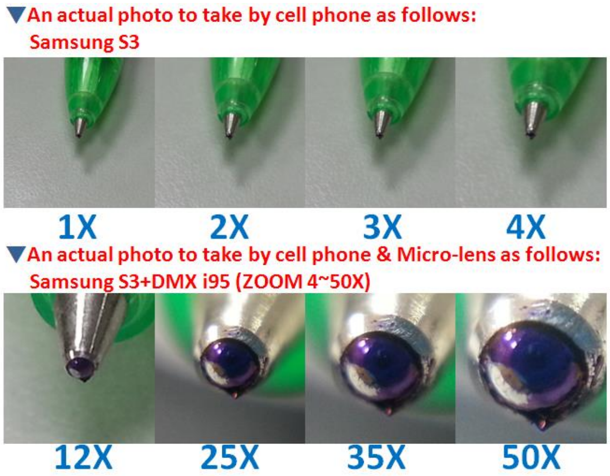
Central Limit Theorem

$$\bar{X}_n \underset{n\to\infty}{\sim} N(\mu, Var(\bar{X}_n) = \sigma^2 / n) \Leftrightarrow \frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} \underset{n\to\infty}{\sim} N(0,1)$$

$LLN: \bar{X}_n - \mu \underset{n\to\infty}{\to} 0$ (convergence to a constant)

$CLT: \sqrt{n}\left(\bar{X}_n - \mu\right) \underset{n\to\infty}{\sim} N(0,\sigma^2)$ (Enlarged version: convergence to a distribution)

Divergence: $n\left(\bar{X}_n - \mu\right) \underset{n\to\infty}{\to} \infty$



Case 1: Normal population $X_i \sim^{iid} N(\mu,\sigma)$, $\dfrac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \underset{\text{any } n}{\sim} N(0,1)$

1. A bottling company uses a filling machine to fill plastic bottles with a popular cola. The bottles are supposed to contain 300 milliliters. In fact, the contents vary according to a normal distribution with mean μ = 298 ml and standard deviation

   = 3 ml. (i.e. $X_i \sim^{iid} N(298, \sigma = 3)$)

   a. What is the probability that <u>an individual bottle</u> contains less than 295 ml?

   $$\Pr(X_i < 295) = \Pr\left(\frac{X_i - \mu}{\sqrt{Var(X_i)}} < \frac{295-298}{3}\right) = \Pr(Z < -1) \quad = 0.1587.$$

   b. What is the probability that the **mean contents** of the bottles in a **six-pack** is less than 295 ml?

   $n = 6$

   $\bar{X} = (X_1 + \ldots + X_6)/6$ → $\bar{X} \sim N(\mu = 298, Var(\bar{X}) = \frac{3^2}{6})$

   $$\Pr(\bar{X} < 295) = \Pr\left(\frac{\bar{X} - \mu}{\sqrt{Var(\bar{X})}} < \frac{295-298}{3/\sqrt{6}}\right)$$

   $$= \Pr(Z < -\sqrt{6}) = \Pr(Z < -2.45) = 0.0071$$

Case 2: Central Limit Theorem

1. X = number of accidents <u>per week</u> = a discrete random variable

$E(X) = 2.2 \quad \sigma = \sqrt{Var(X)} = 1.4$,

Random sample: $(X_1, ..., X_{52})$, $n = 52$, $\bar{X} = \dfrac{X_1 + ... + X_{52}}{52}$

a. What is the approximate distribution of $\bar{X} = \dfrac{X_1 + ... + X_{52}}{52}$?

$E(\bar{X}) = 2.2$, $\sqrt{Var(\bar{X})} = \sigma / \sqrt{n} = 1.4/\sqrt{52} = 0.194$

$\bar{X} \sim^{approximately} N(2.2, \sigma_{\bar{X}} = 0.194)$

b. $\Pr(\bar{X} < 2)$ = the <u>average</u> number of accidents is smaller than 2

$$\Pr(\bar{X} < 2) = \Pr(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{2 - 2.2}{1.4/\sqrt{52}}) \approx \Pr(Z < -1) = 0.1587$$

c. $\Pr(X_1 + ... + X_{52} < 100)$

= the <u>total</u> number of accidents within a year is smaller than 100

Note: You need to convert "total" to "average"

$$\Pr(X_1 + ... + X_{52} < 100) = \Pr(\bar{X} < 100/52) = \Pr(\bar{X} < 1.92)$$

$$\Pr(\bar{X} < 1.92) = \Pr(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{1.92 - 2.2}{1.4/\sqrt{52}}) \approx \Pr(Z < -1.44) = 0.0749$$

Topic: Normal approximation to Binomial distribution (CLT 重要應用)

$Y \sim N(n, p)$, sometimes it is difficult to compute

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

Poisson to approximation for Binomial distribution:

When n large but p small (rare events),

$$\Pr(a \le Y \le b) = \sum_{y=a}^{y=b} \Pr(Y = y) \approx \sum_{y=a}^{y=b} \frac{e^{-np}(np)^y}{y!}$$

Normal approximation for Binomial distribution

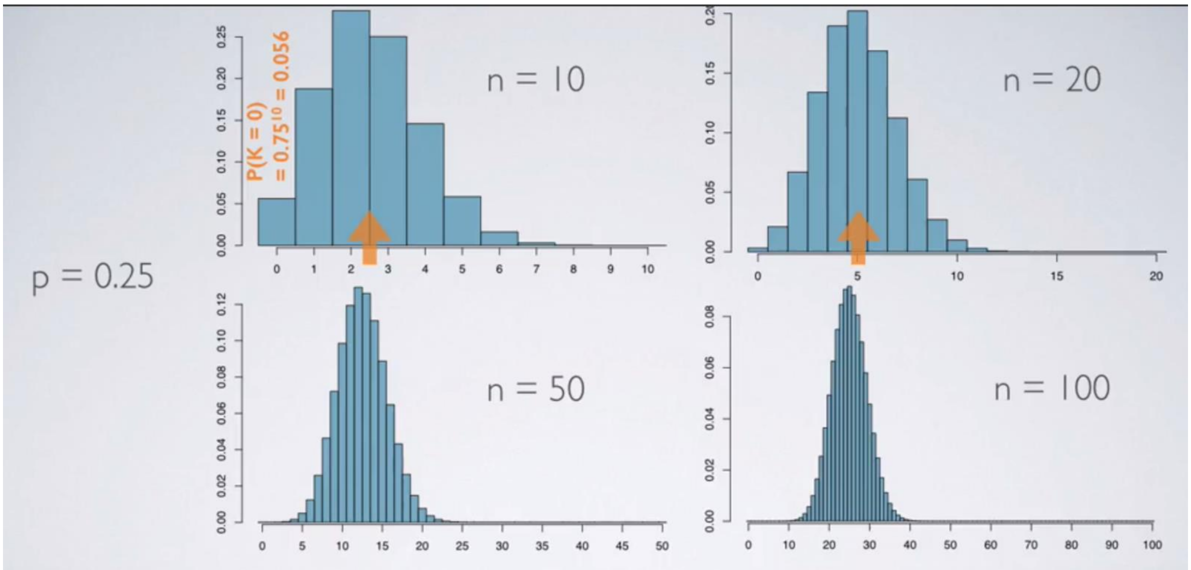
When $n \cdot p \geq 10$ & $n \cdot (1-p) \geq 10$, we can approximate

$$Y \sim N(n, p) \text{ by } X \sim N(\mu = np, \sigma^2 = np(1-p)).$$

Derivations:

$$\begin{aligned}
\Pr(a \leq Y \leq b) &= \Pr(a - np \leq Y - np \leq b - np) \\
&= \Pr\left(\frac{a-np}{\sqrt{np(1-p)}} \leq \frac{Y-np}{\sqrt{np(1-p)}} \leq \frac{b-np}{\sqrt{np(1-p)}}\right) \\
&\approx \Pr\left(\frac{a-np}{\sqrt{np(1-p)}} \leq \frac{X-\mu}{\sigma} \leq \frac{b-np}{\sqrt{np(1-p)}}\right) \quad \text{(by CLT)} \\
&= \Pr\left(\frac{a-np}{\sqrt{np(1-p)}} \leq Z \leq \frac{b-np}{\sqrt{np(1-p)}}\right)
\end{aligned}$$

Plot: fix $p = 0.25$, change n



CLT → distributional property of the sample mean when the same size is large

$$Y = \sum_{i=1}^{n} B_i \quad = \text{total number of successes} = \text{sum of Bernoulli random variables}$$

(i.e. $B_i \sim Bernoulli(p)$, with $E(B_i) = p$ and $Var(B_i) = p(1-p)$)

$$\Pr(a \leq Y \leq b) = \Pr(a \leq \sum_{i=1}^{n} B_i \leq b) = \Pr\left(\frac{a}{n} \leq \frac{\sum_{i=1}^{n} B_i}{n} \leq \frac{b}{n}\right) = \Pr\left(\frac{a}{n} \leq \overline{B} \leq \frac{b}{n}\right)$$

By the Central Limit Theorem,

$$\bar{B} = \hat{p} \sim^{approximately} N(p, Var(\bar{B}) = \frac{Var(B_i)}{n} = \frac{p(1-p)}{n})$$

$$\Pr(\frac{a}{n} \le \hat{p} \le \frac{b}{n}) = \Pr(\frac{\frac{a}{n}-p}{\sqrt{\frac{p(1-p)}{n}}} \le \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \le \frac{\frac{b}{n}-p}{\sqrt{\frac{p(1-p)}{n}}})$$

$$= \Pr(\frac{a-np}{n\sqrt{\frac{p(1-p)}{n}}} \le \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \le \frac{b-np}{n\sqrt{\frac{p(1-p)}{n}}}) \quad \text{(修改)}$$

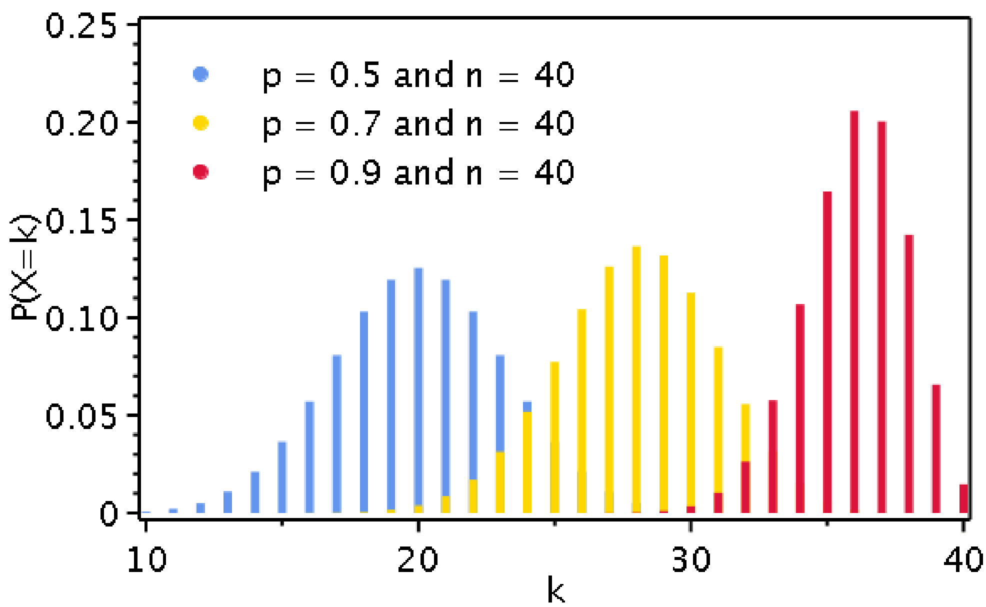
$$\approx \Pr(\frac{a-np}{\sqrt{np(1-p)}} \le Z \le \frac{b-np}{\sqrt{np(1-p)}})$$

Remark: requirement for good approximation → $n \cdot p \ge 10$ **&** $n \cdot (1-p) \ge 10$

When $p = 0.5$, $n \ge 20$; (original Binomial is already symmetric)

When $p = 0.01$, $n \ge 1000$ (original Binomial is very skew)

Plot: fix $n = 40$, change p

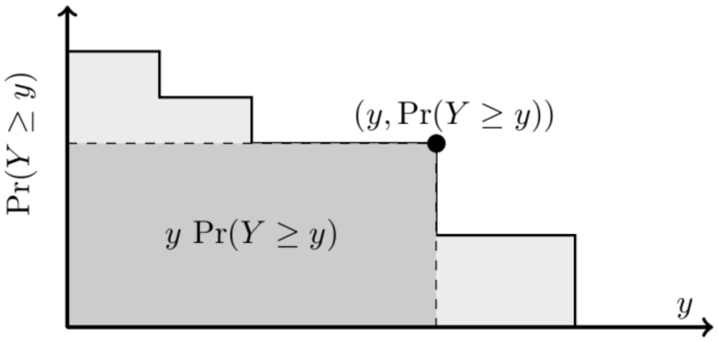


*** History of Normal distribution and Central Limit Theorem**

Markov Inequality – for positive random variables

Let $X > 0$ be a positive random variable.

Markov Inequality: For $a > 0$, $\Pr(X > a) \le \dfrac{E(X)}{a}$.



Let $S(x) = \Pr(X > x) = 1 - \Pr(X \le x) = 1 - F(x)$ = the survival function

For $X > 0$,

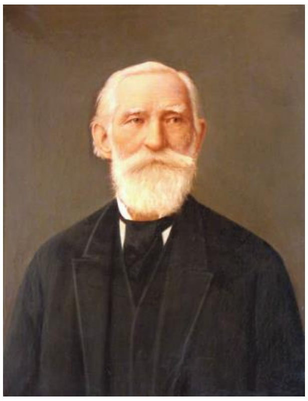
$$
\begin{aligned}
\mu = E(X) &= \int_0^\infty x f(x) dx = -\int_0^\infty x dS(x) \\
&= -\left(xS(x)\big|_0^\infty - \int_0^\infty S(x) dx \right) \\
&= -\{\infty S(\infty) - 0 \cdot S(0)\} + \int_0^\infty S(x) dx \\
&= \int_0^\infty S(x) dx
\end{aligned}
$$

In the graph ($X \to Y, a \to y$), the area under $S(y) = \Pr(Y \ge y)$ is

$$E(Y) = \int_0^\infty S(y) dy = \int_0^\infty \Pr(Y \ge y) dy,$$

which is larger than the area of the rectangle.

Chebyshev (1821-1894) Markov (1852-1922)

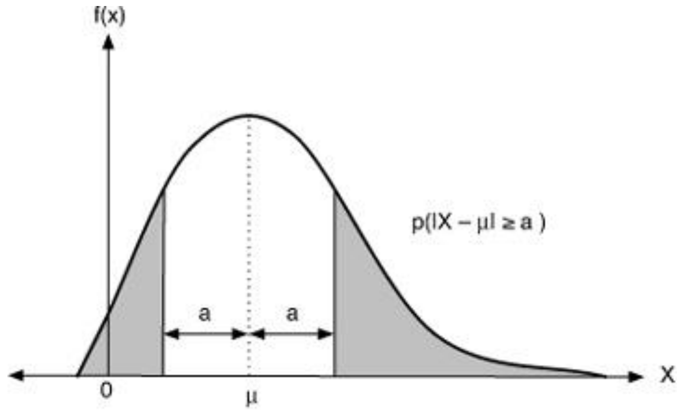


Chebyshev Inequality – for ANY random variables

Version 1: Let X be a random variable with $E(X) = \mu$ and $Var(X) = \sigma^2$.

For $k > 0$

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



Version 2: Let $X_i \sim^{iid}$ $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$

$$\Pr(|\bar{X}-\mu| \geq k) \leq \frac{Var(\bar{X})}{k^2} = \frac{1}{k^2}\frac{\sigma^2}{n}$$ → useful for statistics

Statistical meaning of $\Pr(|\bar{X}-\mu| \geq k)$

- $\bar{X}$ is an estimator of μ
- $|\bar{X}-\mu|$ = estimation error which is a random variable
- $|\bar{X}-\mu| \geq k$ → the error is at least k which is a pre-specified standard

  → a bad thing

Remarks:

- We want the value of error term $|\bar{X}-\mu|$ as small as possible
- Taking the randomness into account, we want $\Pr(|\bar{X}-\mu| \geq k)$ as small as possible.
- Chebyshev's inequality tells you the upper bound of $\Pr(|\bar{X}-\mu| \geq k)$ is $\frac{1}{k^2}\frac{\sigma^2}{n}$.

Example: X = # of items produced in a factory during a week with $E(X) = 50$

Solution: Fact: $X > 0$ → suitable for the Markov inequality

1. $\Pr(X \ge 75) \le \frac{E(X)}{75} = \frac{50}{75}$

The probability that the production exceeds 75 will be no greater than 2/3.

2. $\Pr(40 < X < 60) = ?$

Step 1: Re-express the above two-side inequality in terms of "one-side"

$$\Pr(40 < X < 60) = \Pr(-10 < X - \mu < 10) = \Pr(|X - \mu| < 10)$$

Step 2: Calculate the <u>tail</u> probability first

Based on Chebyshev's inequality

$$\Pr(|X - \mu| \ge 10) = \Pr(|X - \mu|^2 \ge 100) \le \frac{E[|X - \mu|^2]}{100} = \frac{Var(X)}{100} = \frac{25}{100}$$

Step 3: Then calculate the bound for the "within" probability

$$\Pr(|X - \mu| < 10) > 1 - \frac{25}{100} = \frac{3}{4}$$

The probability that the production is between 40 and 60 will be at least 3/4.

Note: $\Pr(|X - \mu| \ge 10) + \Pr(|X - \mu| < 10) = 1$

Remarks on Chebyshev's inequality:

Strength:

It can be applied to any distribution and any sample size → very weak assumption. Hence it has very broad applications.

Weakness: The bound may be too wide and not practical.

Other alternatives to calculate the tail probability

Case 1: $X_i \sim^{iid}$ a known distribution → exact calculation

Case 2: $X_i \sim^{iid}$ unknown distribution

→ apply the central limit theorem for approximation

Example: Chebyshev's approximation vs. exact calculation

(1) Chebyshev vs. Uniform distribution

Given: $E(X)=5$ and $Var(X)=\frac{25}{3}$ → many possibilities

a. Chebyshev's inequality → suitable for any distribution

$$\Pr(|X-5|>4)\le\frac{Var(X)}{4^2}=\frac{1}{16}\cdot\frac{25}{3}=\frac{25}{48}$$

b. Given $X\sim Uniform(0,10)$, $E(X)=\frac{0+10}{2}$

$$Var(X)=\int_0^{10}x^2\frac{1}{10}dx-5^2=\frac{1}{10}\cdot\frac{1}{3}x^3\Big|_0^{10}-25=\frac{100}{3}-\frac{75}{3}=\frac{25}{3}$$

$$\Pr(|X-5|>4)=\Pr(X>9)+\Pr(X<1)=\frac{1}{10}+\frac{1}{10}=\frac{2}{10}$$

Note: $\frac{2}{10}<<\frac{25}{48}$

Conclusion: The bound provided by Chebyshev's inequality is <u>correct but too rough</u>.

(2) Chebyshev vs. normal distribution

Given $E(X)=5$ and $Var(X)=\sigma^2$

a. Chebyshev's inequality

suitable for any distribution with $E(X)=5$, $Var(X)=\sigma^2$

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

b. Normal

$$\Pr(|X-\mu|>2\sigma)=\Pr(\frac{|X-\mu|}{\sigma}>\frac{2\sigma}{\sigma})=\Pr(|Z|>2)=\Pr(Z>2)+\Pr(Z<-2)$$
$$=2\Pr(Z<-2)=0.0456$$

Note: $0.0456<<<\frac{1}{4}$ → The bound is correct but too rough.

Remark: Chebyshev's inequality requires very weak assumption and has very broad applications. However it does not provide precise results.

There is no free lunch!!

Example: The tail probability under normality

Case 1: normal population

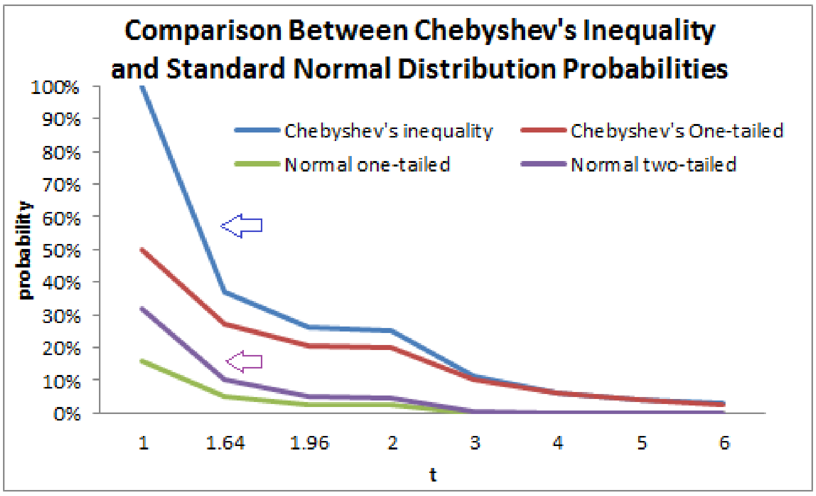
Under normal population: $X_i \sim^{iid} N(\mu,\sigma^2)$:

$$\Pr(|\bar{X}-\mu|>k) \;=\; \Pr(\frac{|\bar{X}-\mu|}{\sigma/\sqrt{n}}>\frac{k}{\sigma/\sqrt{n}})=\Pr(|Z|>\frac{k}{\sigma/\sqrt{n}}) \quad \text{for any n}$$

Case 2: non-normal population but large sample

For any population but with large sample, $X_i \sim^{iid} E(X_i)=\mu,\ Var(X_i)=\sigma^2$

$$\Pr(\frac{|\bar{X}-\mu|}{\sigma/\sqrt{n}}>\frac{k}{\sigma/\sqrt{n}})\approx\Pr(|Z|>\frac{k}{\sigma/\sqrt{n}}) \quad \text{for large n}$$



Implication of Chebyshev's inequality on statistical inference

$$\Pr(|\bar{X}-\mu|\geq k)\leq \frac{Var(\bar{X})}{k^2}=\frac{\sigma^2}{n}\frac{1}{k^2}$$

- $\bar{X}$ is an estimator of μ

- $|\bar{X}-\mu|\geq k \Leftrightarrow$ estimation error is at least k → a bad thing

- $\Pr(|\bar{X}-\mu|\geq k)$ → the smaller, the better

- The probability of the bad thing is bounded by $\frac{\sigma^2}{n}\frac{1}{k^2}$

- Upper bound = $\frac{\sigma^2}{n}\frac{1}{k^2}$

1. k = the standard (人定標準)

   small k → standard is strict

   large k → standard is loose

   $k\uparrow\Rightarrow \frac{\sigma^2}{n}\frac{1}{k^2}\downarrow$

   When you loosen the standard, it becomes less likely to violate it.

2. σ^2 = the variation of the original population (天生)

   $\sigma^2\downarrow\Rightarrow \frac{\sigma^2}{n}\frac{1}{k^2}\downarrow$ (修改箭頭)

   If the original population is less variable, it becomes easier to achieve the standard.

3. n = sample size (後天)

   $n\uparrow\Rightarrow \frac{\sigma^2}{n}\frac{1}{k^2}\downarrow$

   If you get more sample observations, it becomes easier to achieve the standard.

達成目標(降低誤差)的三要素

- 天生麗質 small σ
- 後天努力 large n
- 人定標準寬鬆 k

History of Normal distribution

http://onlinestatbook.com/2/normal_distribution/history_normal.html

Discovery from astronomical observation:

One of the first applications of the normal distribution was to the analysis of errors of measurement made in astronomical observations, errors that occurred because of imperfect instruments and imperfect observers. Galileo (1561-1642, Italian) noted that these errors were symmetric and that small errors occurred more frequently than large errors.

Discovery from coin flip: de Moivre (棣美弗，1667-1754) noted that when the number of events (coin flips) increased, the shape of the binomial distribution approached a very smooth curve.

Formulation: Independently, the mathematicians Adrain in 1808 (USA) and Gauss in 1809 (Germany) developed the formula for the normal distribution and showed that errors were fit well by this distribution. This same distribution had been discovered by Laplace in 1778 (France) when he derived the extremely important *central limit theorem*.

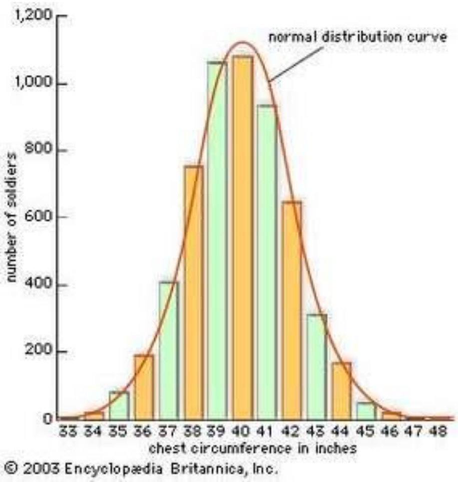
$$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Laplace Gauss Adrain

History of the Central Limit Theorem

The de Moivre–Laplace theorem:

The earliest version (1810) states that the normal distribution may be used as an approximation to the binomial distribution.



Towards the end of the 19th century, the mathematical discussion was turning increasingly from computational mathematics to a more fundamental analysis, to "pure" mathematics. This had a big impact on probability theory as it had been considered more as "common sense" than a rigorous mathematical theory.

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.530.4669&rep=rep1&type=pdf

Several Versions of the Central Limit Theorem

- *Lindeberg–Lévy CLT:*

$\{X_1,...,X_n\}$ is a sequence of *iid* sample with $E(X_i)=\mu$ and $Var(X_i)=\sigma^2<\infty$

- *Lyapunov CLT:* Suppose $\{X_1,...,X_n\}$ is a sequence of *independent* sample with $E(X_i)=\mu_i$ and $Var(X_i)=\sigma_i^2<\infty$. Lyapunov derived a condition to show that

$$\left(\sum_{i=1}^{n}\sigma_i^2\right)^{-1/2}\sum_{i=1}^{n}(X_i-\mu_i)$$

- *Lindeberg CLT:* Lindeberg (1920) derived a weaker condition (stronger result) for the theorem. Lindeberg's work was unknown to Alan Turing, who proved the central limit theorem in his dissertation in 1935.

Subsequent work:

The classical central limit theorem assumes identically distributed random variables with finite variance. The Lindeberg- Lévy -Feller central limit theorem showed that we can weaken the condition of identically distributed random variables so long as they satisfy the Lindeberg condition. It is natural to ask what happens if instead of weakening the identically distributed hypothesis, we weaken dependence.