Limit Theorem
极限定理

Probability and Mathematical Statistics

(概率与数理统计)

Xi ZHANG

In many cases, we don't need to calculate exactly the probability but roughly know it

- Especially when the probability is very large or very small
  - e.g P{haze tomorrow in Tahiti} = ?

- For example, suppose tossing a coin 1000 times, we would like to know if the probability of consecutive 17 appearance of heads
  - Let N be the number of occurrences of 17 consecutive heads in 1000 coin flips.

$$N = I_1 + \ldots + I_{984}$$

$$E[I_i] = P(I_i = 1) = 1/2^{17}$$

$$E[N] = 984 \cdot 1/2^{17} = 0.007507$$

Outlines

- Chebyshev's Inequality and the Weak Law of Large Numbers (切比雪夫不等式及弱大数定律）
- The Central Limit Theorem(中心极限定理)
- The Strong Law of Large Numbers（强大数定律）
- Summary

Markov's Inequality （马尔可夫不等式）

Proposition: Markov's Inequality

If X is a random variable that takes only nonnegative values, then, for any value $a > 0$

$$P\{X \geq a\} \leq \frac{E[X]}{a}$$

Hence, $P[N \geq 1] \leq E[N] \,/\, 1 \leq 0.75\%.$

$$E[X] = E[X \mid X \geq a]\, P(X \geq a) + E[X \mid X < a]\, P(X < a)$$

$$\geq a \qquad \geq 0 \qquad \geq 0$$

$$E[X] \geq a\, P(X \geq a) + 0.$$

Chebyshev's Equality (切比雪夫不等式)

Proposition: Chebyshev's Inequality

X is a random variable with finite mean μ and variance σ^2, then, for any value $k > 0$,

$$P\{|X-\mu| \geq k\} \leq \frac{\sigma^2}{k^2}$$

$$P(|X-\mu| \geq k) = P((X-\mu)^2 \geq k^2) \leq E[(X-\mu)^2] / k^2 = \sigma^2 / k^2$$

Examples

- Suppose that it is known that the number of items produced in a factory during a week is a random variable with mean 50.

  (a) What can be said about the probability that this week's production will exceed 75?

  (b) If the variance of a week's production is known to equal 25, then what can be said about the probability that this week's production will be between 40 and 60?

- Exercise
  - If X is uniformly distributed over the interval (0, 10), What can be said about the probability $P\{|X-5|>4\}$

The Weak Law of Large Numbers （弱大数定理）

Proposition:

If $Var(X) = 0$, then

$$P\{X = E[X]\} = 1$$

Theorem: The weak law of large numbers

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables, each having finite mean $E[X_i] = \mu$. Then, for any $\varepsilon > 0$,

$$P\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \varepsilon\right\} \to 0 \quad \text{as } n \to \infty$$

The Central Limit Theorem （中心极限定理）

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables, each having mean μ and variance σ^2. Then the distribution of

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

tends to the standard normal as $n \to \infty$. That is, for $-\infty < a < \infty$,

$$P\left\{\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \le a\right\} \to \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a} e^{-x^2/2}dx \text{ as } n \to \infty$$

The Central Limit Theorem （中心极限定理）

Central limit theorem for independent random variables

Let $X_1, X_2, \ldots$ be a sequence of independent random variables having respective means and variances $\mu_i = E[X_i], \sigma_i^2 = \mathrm{Var}(X_i)$. If (a) the X_i are uniformly bounded—that is, if for some M, $P\{|X_i| < M\} = 1$ for all i, and (b) $\sum_{i=1}^{\infty} \sigma_i^2 = \infty$—then

$$P\left\{\frac{\sum_{i=1}^{n}(X_i - \mu_i)}{\sqrt{\sum_{i=1}^{n}\sigma_i^2}} \le a\right\} \to \Phi(a) \quad as \quad n \to \infty$$

Example

- An instructor has 50 exams that will be graded in sequence. The times required to grade the 50 exams are independent, with a common distribution that has mean 20 minutes and standard deviation 4 minutes. Approximate the probability that the instructor will grade at least 25 of the exams in the first 450 minutes of work.

$$X = \sum_{i=1}^{25} X_i$$

$$E[X] = \sum_{i=1}^{25} E[X_i] = 25 \times 20 = 500$$

$$Var(X) = \sum_{i=1}^{25} Var[X_i] = 25 \times 16 = 400$$

$$P\{X \leq 450\} = P\left\{\frac{X - 500}{\sqrt{400}} \leq \frac{450 - 500}{\sqrt{400}}\right\} \approx P\{Z \leq -2.5\} = 1 - \Phi(2.5) = 0.006$$

Exercise

If 10 fair dice are rolled, find the approximate probability that the sum obtained is between 30 and 40, inclusive.

The Strong Law of Large Numbers

Theorem: The strong law of large numbers

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed ra ndom variables, each having a finite mean $\mu = E[X_i]$. Then, with proba bility 1,

$$\frac{X_1 + X_2 + \cdots + X_n}{n} \to \mu \quad \text{as } n \to \infty$$

Comparison between weak law of large numbers

- Weak law of large numbers: For any specified large value n^*, $\frac{X_1 + \cdots + X_{n^*}}{n^*}$ is likely to be near μ, it does not say that $(X_1 + \cdots + X_n)/n$ is bound to stay near μ for all values of n larger than n^*. Thus, it leaves open the possibility that large valu es of $|(X_1 + \cdots + X_n)/n - \mu|$ can occur infinitely often (though at infrequen t intervals).
- The strong law shows that this cannot occur

Summary

- Chebyshev's Equality $$P\{|X-\mu| \geq k\} \leq \frac{\sigma^2}{k^2}$$
- Weak Law of Large Numbers $$P\left\{\left|\frac{X_1+\cdots+X_n}{n}-\mu\right| \geq \varepsilon\right\} \leq 0 \quad \text{as } n \to \infty$$
- The Central Limit Theorem $$P\left\{\frac{X_1+\cdots+X_n-n\mu}{\sigma\sqrt{n}} \leq a\right\} \to \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a} e^{-x^2/2}dx \text{ as } n \to \infty$$
- The Strong Law of Large Numbers $$\frac{X_1+X_2+\cdots+X_n}{n} \to \mu \quad \text{as } n \to \infty$$

Markov chains@stochastic processes

- Stochastic processes
  - Many real-world systems contain uncertainty and evolve over time.
  - Stochastic processes (and Markov chains) are probability models for such systems.
  - A discrete-time stochastic process is a sequence of random variables:

$$X_0, X_1, X_2, \ldots \text{ typically denoted by } \{X_t\}$$

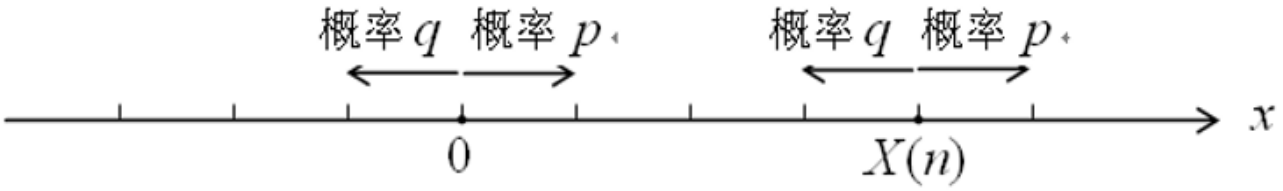
Time: $t = 0, 1, 2, \ldots$

State: v-dimensional vector, $\boldsymbol{s} = (s_1, s_2, \ldots, s_v)$

In general, there are m states (a finite # of states): $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_m$

Random walk problem

A stochastic process whose state space is given by the integer $i = 0, \pm 1, \pm 2, \ldots$ is said to be a random walk if, for some number $0 < p < 1$, $P_{i,i+1} = p = 1 - P_{i,i-1}, i = 0, \pm 1, \pm 2, \ldots$



A Markov Chain

Definition of the Markov chain

- A stochastic process $\{X_t\}$ is called a Markov chain if

$$P\{X_{t+1} = j | X_t = i, X_{t-1} = i_{t-1}, \dots, X_1 = i_1, X_0 = i_0\}$$
$$= P\{X_{t+1} = j | X_t = i\} = P_{ij}$$

← transition probabilities
转移概率

Discrete time means $t \quad T = \{0, 1, 2, \dots\}$

The future behavior of the system depends only on the current state i and not on any of the previous states

For all M states, $\sum_j p_{ij} = 1 \quad i \quad and \quad p_{ij} \geq 0 \quad i, j$

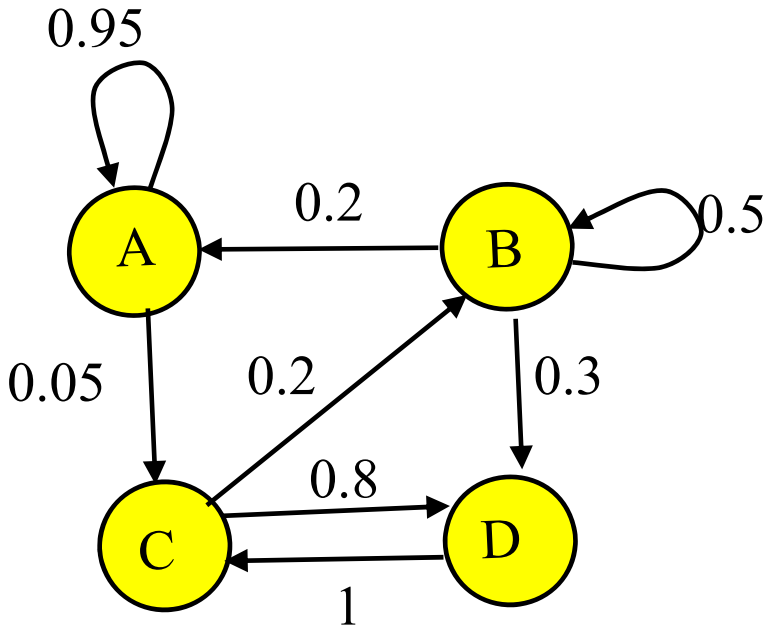
$$\begin{bmatrix} P_{00} & \cdots & P_{0M} \\ \vdots & \ddots & \vdots \\ P_{M0} & \cdots & P_{MM} \end{bmatrix}$$

← transition matrix
转移概率矩阵

$$P\{X_t = i, X_{t-1} = i_{t-1}, \dots, X_1 = i_1, X_0 = i_0\} = P_{i_{t-1}, i_t} P\{X_{t-1} = i_{t-1}, \dots, X_1 = i_1, X_0 = i_0\}$$
$$= P_{i_{t-1}, i_t} P_{i_{t-2}, i_{t-1}} \dots P_{i_1, i_2} P_{i_0, i_1} P\{X_0 = i_0\}$$

Matrix representation

	A	B	C	D
A	0.95	0	0.05	0
B	0.2	0.5	0	0.3
C	0	0.2	0	0.8
D	0	0	1	0



Each directed edge A→B is associated with the **positive** transition probability from A to B

Example_Gambler's Ruin

At time zero the gambler has X_0 = \$2, and each day he makes a \$1 bet. He wins with probability p and lose with probability $1-p$. He will quit if he ever obtains \$4 or if loses all my money.

X_t = amount of money he has after the bet on day t. State space is $S = \{0, 1, 2, 3, 4\}$

So, $X_1 = \begin{cases} 3 \text{ with probability } p \\ 1 \text{ with probability } 1-p \end{cases}$

if $X_t = 4$ then $X_{t+1} = X_{t+2} = \bullet\bullet\bullet = 4$, and
if $X_t = 0$ then $X_{t+1} = X_{t+2} = \bullet\bullet\bullet = 0$.

	0	1	2	3	4
0	1	0	0	0	0
1	$1-p$	0	p	0	0
2	0	$1-p$	0	p	0
3	0	0	$1-p$	0	p
4	0	0	0	0	1

n-step Transition Probabilities

- Let p_{ij} be probability of going from i to j in two transitions
  - In matrix form, $\mathbf{P}^{(2)} = \mathbf{P} \times \mathbf{P}$
  - For $n = 3$: $\mathrm{P}^{(3)} = \mathrm{P}^{(2)}\,\mathrm{P} = \mathrm{P}^2\,\mathrm{P} = \mathrm{P}^3$
    - more generally, $\mathrm{P}^{(n)} = \mathrm{P}^{(m)}\ \mathrm{P}^{(n-m)}$

The ij th entry of this reduces to

$$p_{ij}^{(n)} = \sum_{k=0}^{M} p_{ik}^{(m)}\, p_{kj}^{(n-m)} \qquad 1 \le m \le n-1$$

Chapman - Kolmogorov Equations

Revisit Gambler's Ruin problem

- Sometimes, the components P_{ij}^{n} in the transition matrix will converge
  - Limiting probability (极限概率) π_j, no matter what the initial state was.
- Gambler's Ruin with $p = 0.75, t = 30$

	0	1	2	3	4
0	1	0	0	0	0
1	0.25	0	0.75	0	0
2	0	0.25	0	0.75	0
3	0	0	0.25	0	0.75
4	0	0	0	0	1

$\mathbf{P}^{(30)} =$

	0	1	2	3	4
0	1	0	0	0	0
1	0.325	ε	0	ε	0.675
2	0.1	0	ε	0	0.9
3	0.025	ε	0	ε	0.975
4	0	0	0	0	1

Probabilities

n-Step Transition Matrix

30 step Transition Matrix

		0	1	2	3	4
		"$0"	"$1"	"$2"	"$3"	"$4"
0	"$0"	1	0	0	0	0
1	"$1"	0.325	2.04E-07	0.00E+00	6.12E-07	0.674999
2	"$2"	0.1	0.00E+00	4.08E-07	0.00E+00	0.9
3	"$3"	0.025	6.80E-08	0.00E+00	2.04E-07	0.975
4	"$4"	0	0	0	0	1

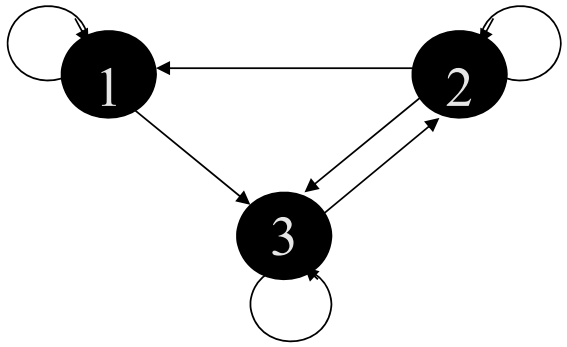
Steady-State Probabilities

Let $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_m)$ is the m-dimensional row vector of steady-state (unconditional) probabilities for the state space $\boldsymbol{S} = \{1,\ldots,m\}$. To find steady-state probabilities, solve linear system:

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P},\ \sum_{j=1,m} \pi_j = 1,\ \pi_j \geq 0,\ j = 1,\ldots,m$$

Steady-state probabilities might not exist unless the Markov chain is ergodic （遍历的）

A Markov chain is ergodic if it is aperiodic and allows the attainment of any future state from any initial state after one or more transitions.

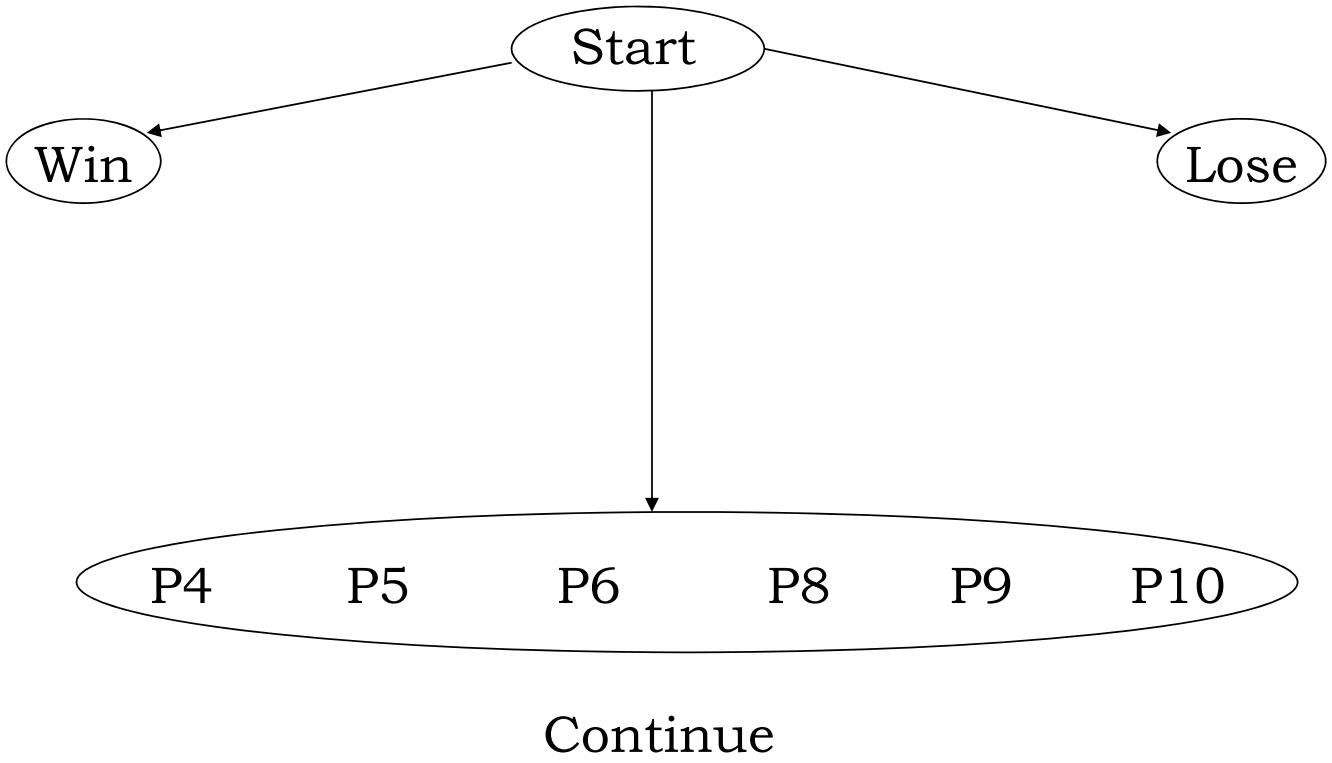


Game of Craps

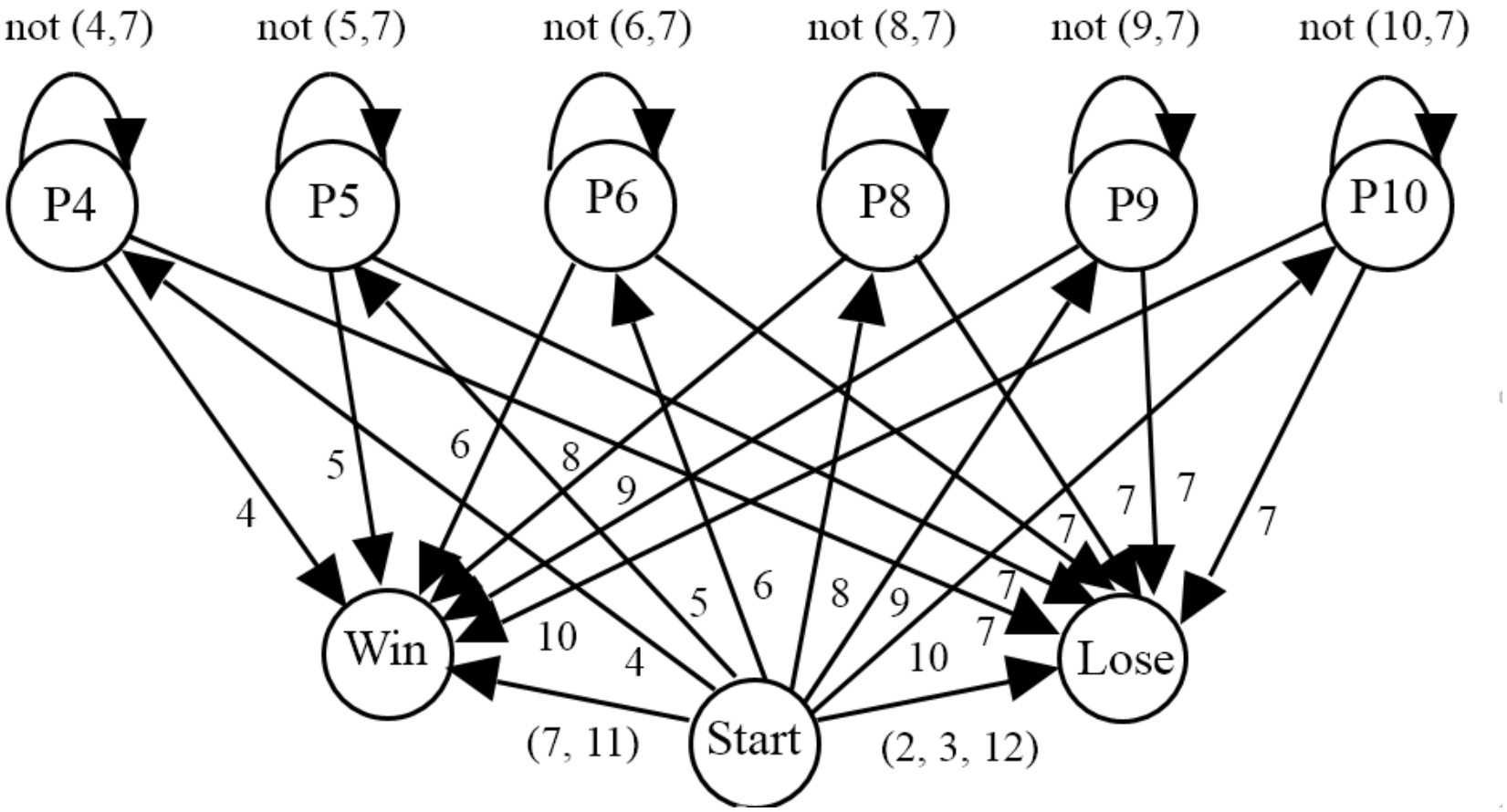
- The rules of craps is as follows:
  - The player rolls a pair of dice and sums the numbers showing.
    - A total of 7 or 11 on the first rolls wins for the player
    - Where a total of 2, 3, 12 loses
    - Any other number is called the point.
  - The player rolls the dice again.
    - If she rolls the point number, she wins
    - If she rolls number 7, she loses
    - Any other number requires another roll
  - The game continues until he/she wins or loses

Game of Craps as a Markov Chain

All the possible states



Game of Craps Network



Probability in craps

Sum	2	3	4	5	6	7	8	9	10	11	12
Prob.	0.028	0.056	0.083	0.111	0.139	0.167	0.139	0.111	0.083	0.056	0.028

e.g.

- Probability of win = $P\{7\ or\ 11\} = 0.167 + 0.056 = 0.223$
- Probability of loss = $P\{2, 3, 12\} = 0.028 + 0.56 + 0.028 = 0.112$

P =

	Start	Win	Lose	P4	P5	P6	P8	P9	P10
Start	0	0.222	0.111	0.083	0.111	0.139	0.139	0.111	0.083
Win	0	1	0	0	0	0	0	0	0
Lose	0	0	1	0	0	0	0	0	0
P4	0	0.083	0.167	0.75	0	0	0	0	0
P5	0	0.111	0.167	0	0.722	0	0	0	0
P6	0	0.139	0.167	0	0	0.694	0	0	0
P8	0	0.139	0.167	0	0	0	0.694	0	0
P9	0	0.111	0.167	0	0	0	0	0.722	0
P10	0	0.083	0.167	0	0	0	0	0	0.75

Transient Probabilities $\mathbf{q}^{(n)}$ in Craps

Roll no.	Start	Win	Lose	P4	P5	P6	P8	P9	P10
0	1	0	0	0	0	0	0	0	0
1	0	0.222	0.111	0.083	0.111	0.139	0.139	0.111	0.083
2	0	0.299	0.222	0.063	0.08	0.096	0.096	0.080	0.063
3	0	0.354	0.302	0.047	0.058	0.067	0.067	0.058	0.047
4	0	0.394	0.359	0.035	0.042	0.047	0.047	0.042	0.035
5	0	0.422	0.400	0.026	0.030	0.032	0.032	0.030	0.026

This is not an ergodic Markov chain so where you start is important

Absorbing State Probabilities for Craps

Initial state	Win	Lose
Start	0.493	0.507
P4	0.333	0.667
P5	0.400	0.600
P6	0.455	0.545
P8	0.455	0.545
P9	0.400	0.600
P10	0.333	0.667

Interpretation of Steady-State Conditions

- Just because an ergodic system has steady-state probabilities does not mean that the system "settles down" into any one state
- π_j is simply the likelihood of finding the system in state *j* after a large number of steps
- The limiting probability π_j that the process is in state *j* after a large number of steps is also equals the long-run proportion of time that the process will be in state *j*

Point Estimation

Probability and Mathematical Statistics

(概率与数理统计)

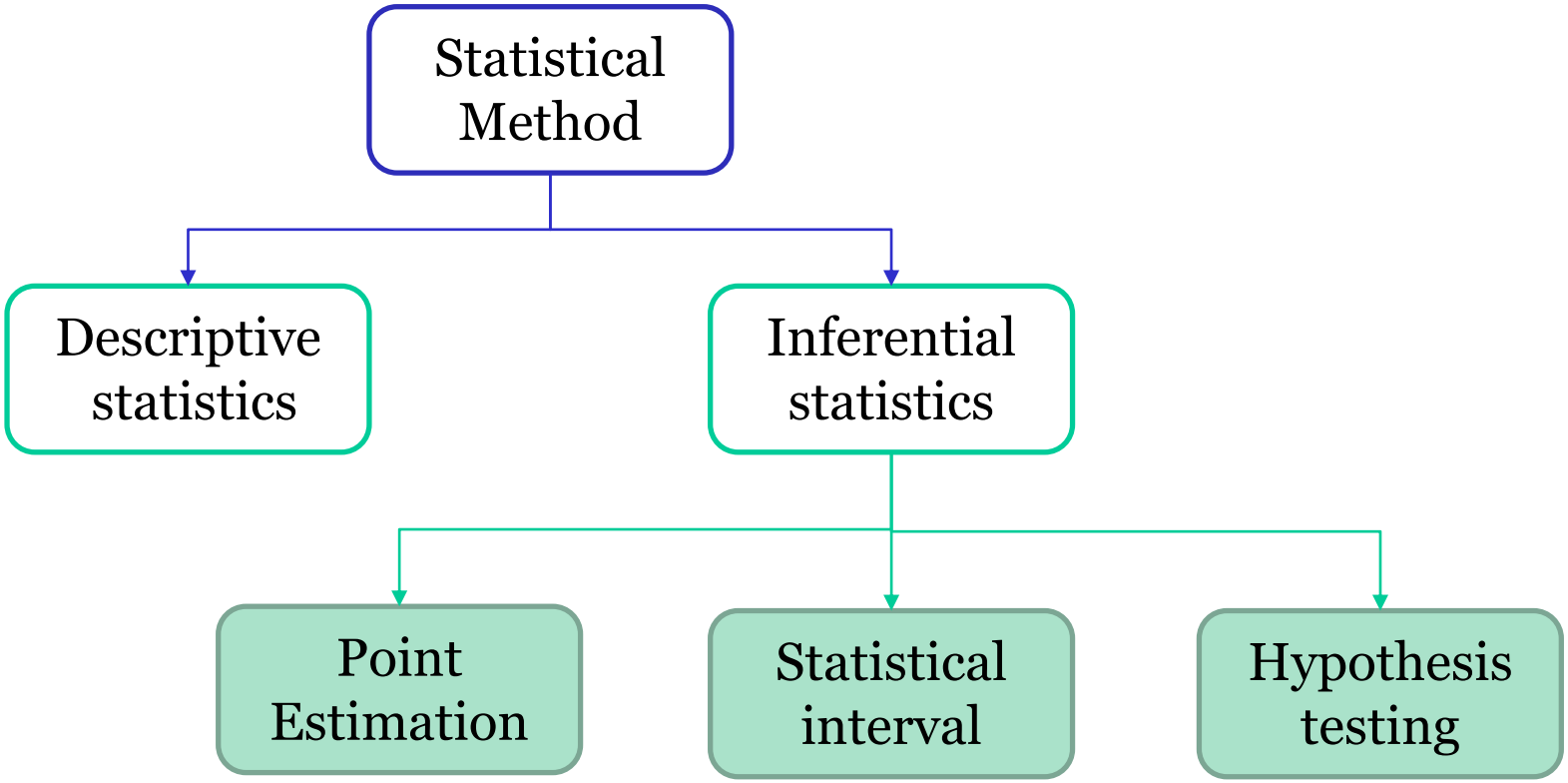
Xi ZHANG

College of Engineering

Outline

- Descriptive statistics
- Point estimation
  - Moment estimation
  - Maximum likelihood estimation
- Evaluating the estimators
- Sampling distributions

Statistical methods



Examples

某工厂生产大批的电子元件，在实际应用中，我们可以提出许多感兴趣的问题，

例如：（1）元件的平均寿命如何？

参数估计

（2）有理由假设元件的寿命服从指数分布，该假设符合实际情况吗？

假设检验

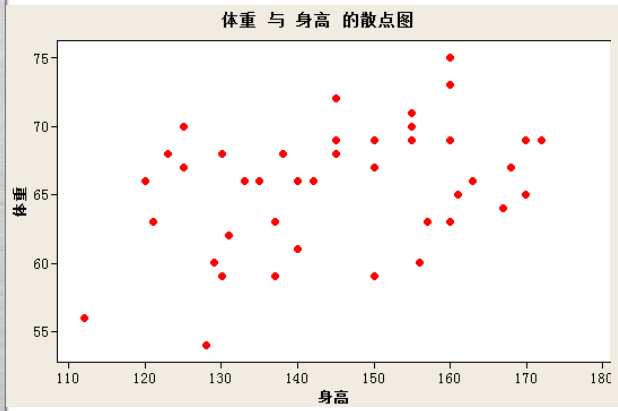
国家质检总局第一次全国液态奶三聚氰胺专项检查的结果：

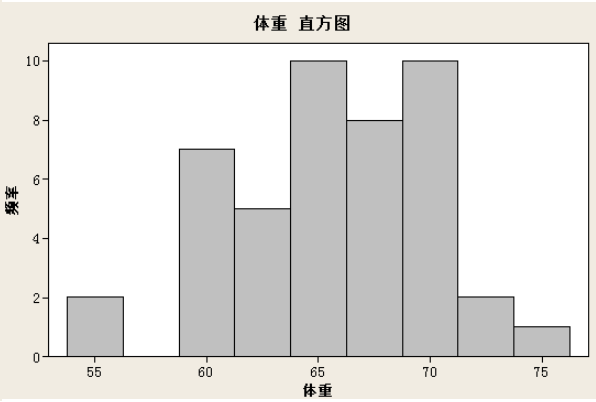
- 抽检蒙牛产品121批次，11批次检出三聚氰胺；
- 抽检伊利产品81批次，7批次检出三聚氰胺；
- 抽检三元产品53批次，53批次均未检出三聚氰胺；
- 抽检雀巢产品7批次，7批次均未检出三聚氰胺。

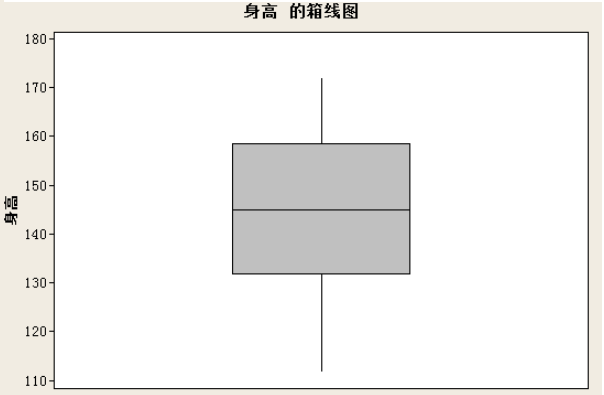
哪个品牌更值得信赖？

Descriptive statistics

- Target : quantitatively describing the main features of a collection of information
  - e.g. : central tendency, shape, *etc.*
  - Tools we can use:
    - Table
    - Graphs
      - Scatter plot（散点图）
      - Histogram（直方图）
      - Box plot（箱线图）
      - Probability plot
      - *etc.*



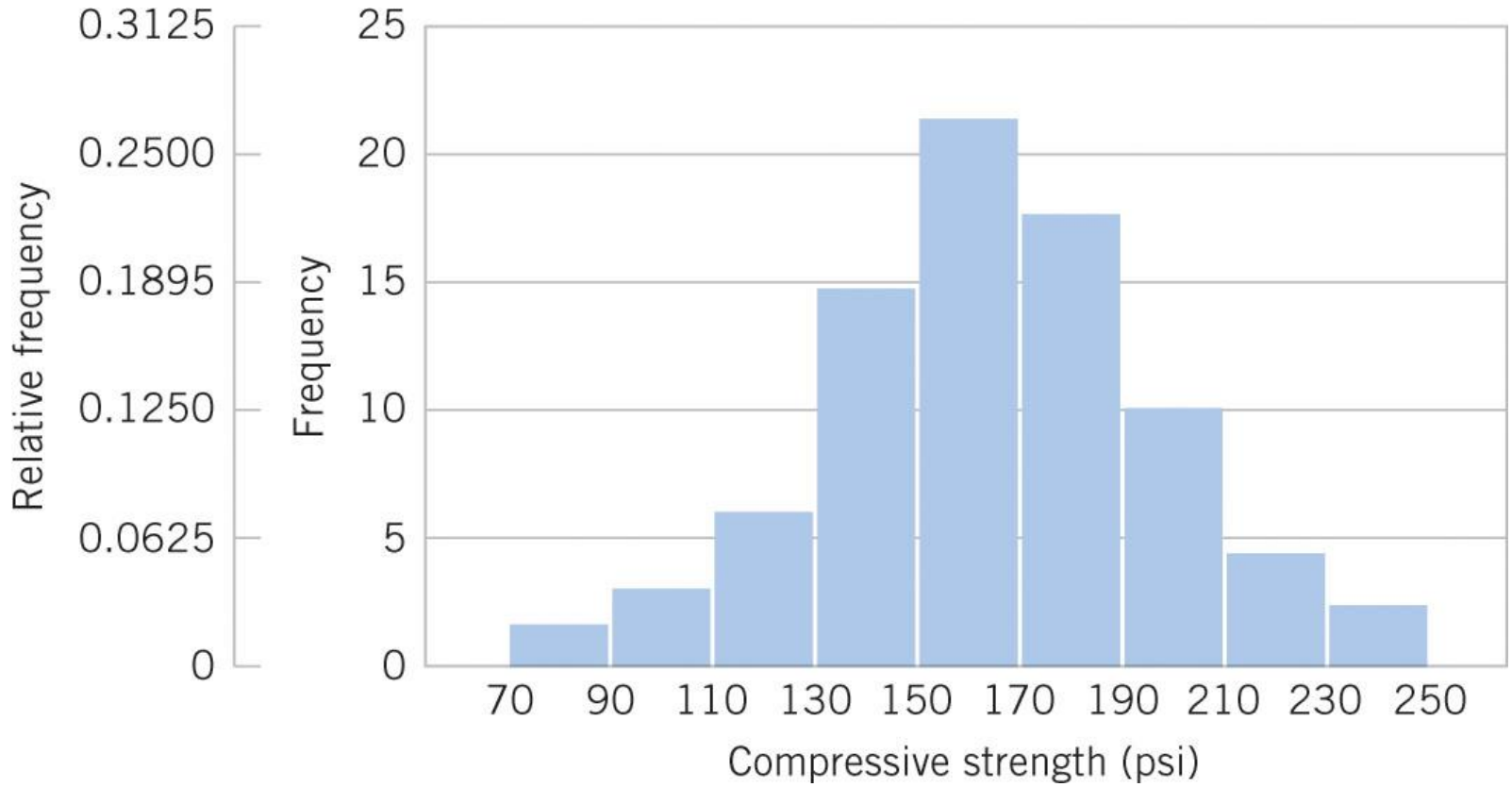




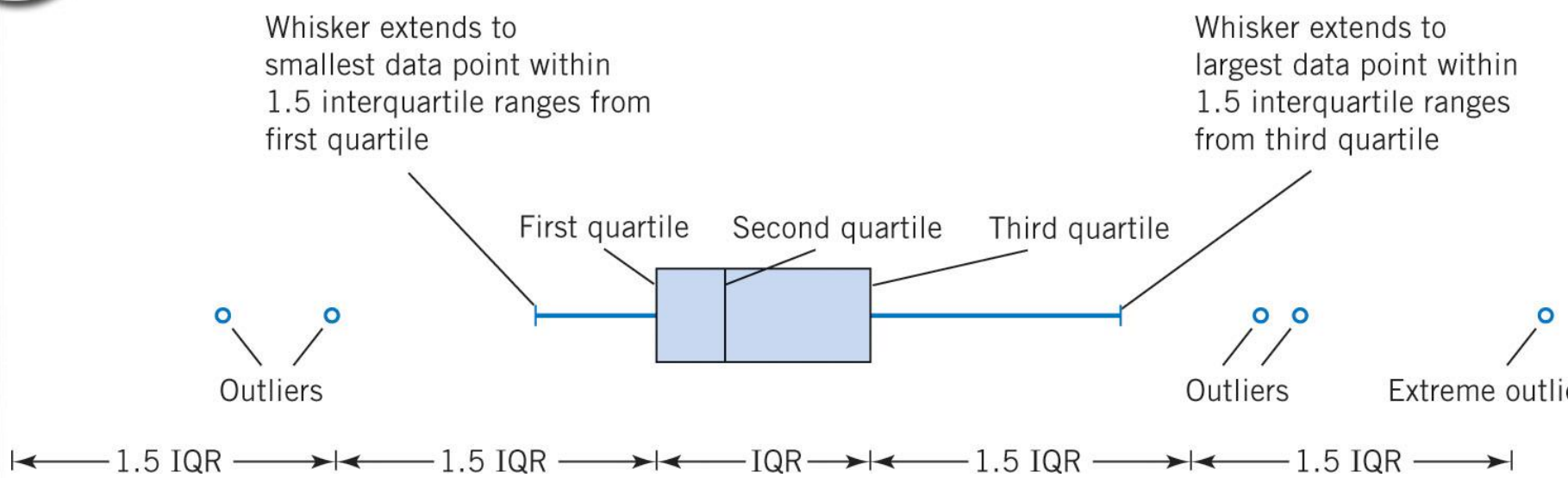
Histogram

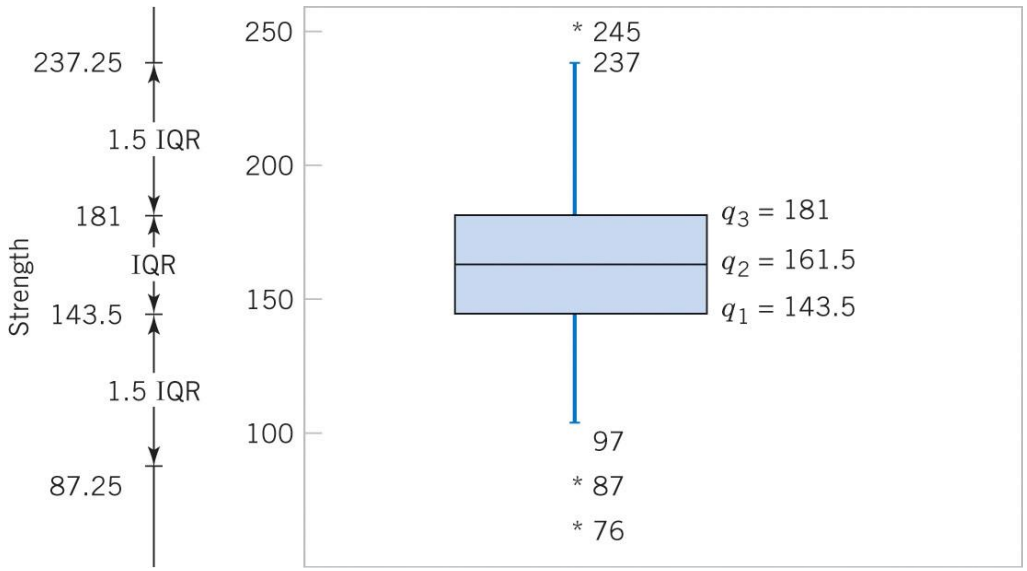
- 采集到某汽车零件的抗压强度

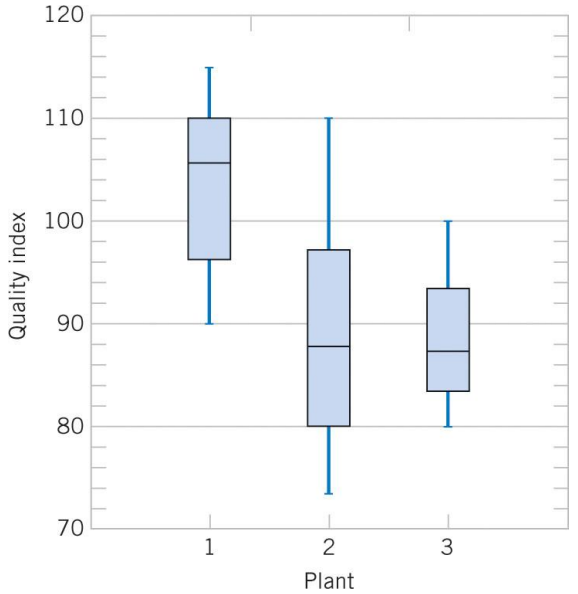
Class	$70 \le x < 90$	$90 \le x < 110$	$110 \le x < 130$	$130 \le x < 150$	$150 \le x < 170$	$170 \le x < 190$	$190 \le x < 210$	$210 \le x < 230$	$230 \le x < 250$
Frequency	2	3	6	14	22	17	10	4	2
Relative frequency	0.0250	0.0375	0.0750	0.1750	0.2750	0.2125	0.1250	0.0500	0.0250
Cumulative relative frequency	0.0250	0.0625	0.1375	0.3125	0.5875	0.8000	0.9250	0.9750	1.0000



Box plot


Whisker extends to smallest data point within 1.5 interquartile ranges from first quartile
Whisker extends to largest data point within 1.5 interquartile ranges from third quartile
First quartile
Second quartile
Third quartile
Outliers
Outliers
Extreme outli
1.5 IQR
1.5 IQR
IQR
1.5 IQR
1.5 IQR

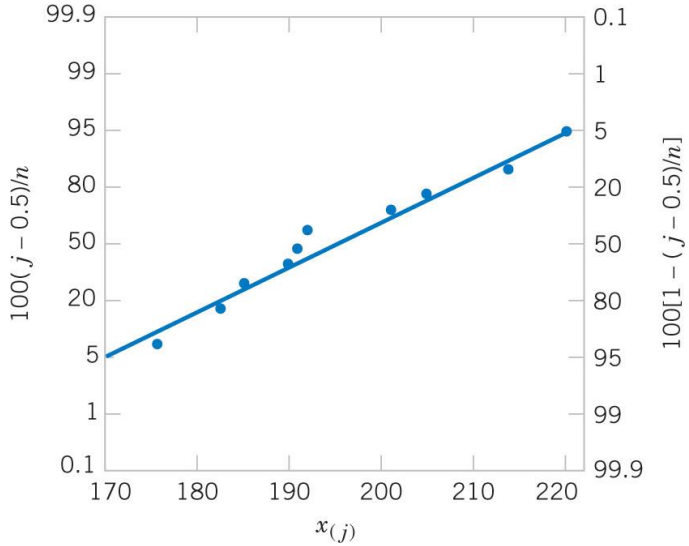

Strength
237.25
1.5 IQR
181
IQR
143.5
1.5 IQR
87.25
250
200
150
100
* 245
237
q_3 = 181
q_2 = 161.5
q_1 = 143.5
97
* 87
* 76

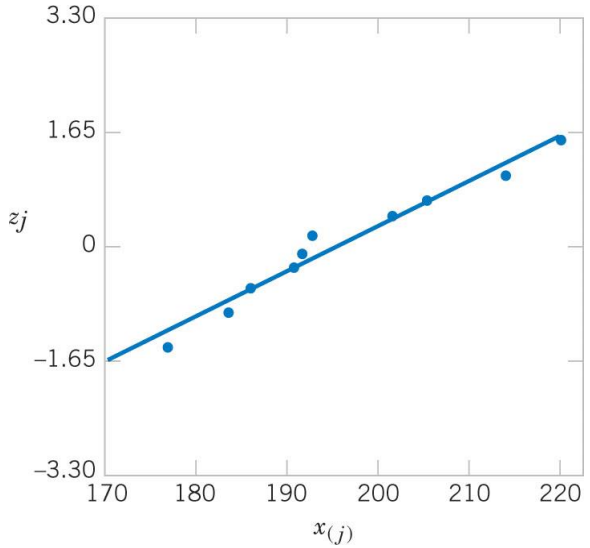

Quality index
120
110
100
90
80
70
1
2
3
Plant

Probability plot

- It is a graphical method for determining whether sample dat a conform to a hypothesized distribution based on a subjecti ve visual examination of the data
  - Vertical axes are scaled to specified distribution we want to test

j	$x_{(j)}$	$(j - 0.5)/10$	z_j
1	176	0.05	−1.64
2	183	0.15	−1.04
3	185	0.25	−0.67
4	190	0.35	−0.39
5	191	0.45	−0.13
6	192	0.55	0.13
7	201	0.65	0.39
8	205	0.75	0.67
9	214	0.85	1.04
10	220	0.95	1.64





Point Estimation （点估计）

- A single number calculated from sample data for which we have some expectation, or assurance, that is reasonably close the par ameter it is supposed to estimate
- Point estimator
  - Any function $\hat{\Theta} = h(X_1, X_2, \ldots X_n,)$ of a sample to estimate populat ion parameter
  - e.g. Sample mean $\bar{x} = 7$
- Example: Engineers would like to detect the strength of a kind of alloy steel, and 12 experiments are conducted for data collection as follows

1	2	3	4	5	6	7	8	9	10	11	12
2.4	2.9	2.7	2.6	2.9	2.0	2.8	2.2	2.4	2.0	2.4	2.5

$$\bar{x} = \frac{29.8}{12} = 2.483 \qquad s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} = \frac{75.08 - (29.8)^2/12}{12-1} = 0.09788$$

Moment estimation （矩估计）

- Let $X_1, X_2, ..., X_n$ be a sample from a population with pdf/pmf $f(x|\theta_1, \theta_2, ..., \theta_k)$. Method of moments estimators are found by equating the first ***k*** sample moments to the corresponding *k* population moments, and solving the resulting system of sim ultaneous equations

Let $X_1, X_2, \cdots, X_n$ be IID samples of a random variable X. The kth sample moment is (样本矩)

$$m_k := \bar{X_n^k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i, m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2, \cdots, m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

$$u'_1 = E(X), u'_2 = E(X^2), ..., u'_k = E(X^k)$$

Example (Normal distribution moment estimators)

- Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution with parameters μ and σ^2.

$$E(X) = \mu \text{ and } E(X^2) = \mu^2 + \sigma^2.$$

$$\mu = \bar{X}, \qquad \mu^2 + \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

$$\hat{\mu} = \bar{X}, \qquad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2}{n} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n}$$

Exercise

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a binomial distribution with parameters k and p.

Maximum likelihood estimation （极大似然估计）

- If $X_1, X_2, \ldots, X_n$ are i.i.d. samples from a population with pdf/pmf $f(x|\theta_1, \theta_2, \ldots, \theta_k)$, defining the likelihood function （似然函数）of the sample

$$L(\theta \mid x) = L(\theta_1, \theta_2, \ldots, \theta_k \mid x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i \mid \theta_1, \theta_2, \ldots, \theta_k)$$

  - Then the maximum likelihood estimator of θ is the value of θ that maximizes the likelihood function L(θ)

- Example Bernoulli distribution

$$f(x; p) = \begin{cases} p^x (1-p)^{1-x}, & x = 0, 1 \\ 0, & \text{otherwise} \end{cases}$$

$$L(p) = p^{x_1}(1-p)^{1-x_1} p^{x_2}(1-p)^{1-x_2} \cdots p^{x_n}(1-p)^{1-x_n}$$

$$= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$$

$$\ln L(p) = \left(\sum_{i=1}^{n} x_i\right) \ln p + \left(n - \sum_{i=1}^{n} x_i\right) \ln(1-p)$$

$$\frac{d \ln L(p)}{dp} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{\left(n - \sum_{i=1}^{n} x_i\right)}{1-p}$$

- Exercise： normal distribution

Bayesian method

- θ is considered to be a quantity whose variation can be described by a probability distribution(prior distribution先验分布). A sample is taken from a population indexed by θ and the prior distribution is updated with the sample information (posterior 后验分布)
- Bayesian estimator
  - Denote prior distribution by π(θ) and sampling distribution by $f(x|\theta)$, the posterior is

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{m(x)}$$

$$m(x) = \int f(x \mid \theta)\pi(\theta)d\theta$$

Bayesian method

- Example Bayes estimator for the mean of a Normal Dis tribution (μ is unknown, and assume the prior distribu tion for μ is normal with mean μ_0 and variance σ_0^2)

$$f(\mu) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-(\mu-\mu_0)^2/(2\sigma_0^2)} = \frac{1}{\sqrt{2\pi}\sigma_0^2} e^{-(\mu^2-2\mu_0\mu+\mu_0^2)/(2\sigma_0^2)}$$

The joint probability distribution of the sample is

$$f(x_1, x_2, \ldots, x_n \mid \mu) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2\sigma^2)\sum_{i=1}^{n}(x_i-\mu)^2}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2\sigma^2)(\sum x_i^2 - 2\mu\sum x_i + n\mu^2)}$$

The joint probability distribution of the sample and μ is

$$f(x_1, x_2, \ldots, x_n, \mu) = \frac{1}{(2\pi\sigma^2)^{n/2}\sqrt{2\pi}\sigma_0} \times e^{-(1/2)[(1/\sigma_0^2+n/\sigma^2)\mu^2-(2\mu_0/\sigma_0^2+2\sum x_i/\sigma^2)\mu+\sum x_i^2/\sigma^2+\mu_0^2/\sigma_0^2]}$$

$$= e^{-(1/2)\left[\left(\frac{1}{\sigma_0^2}+\frac{1}{\sigma^2/n}\right)\mu^2-2\left(\frac{\mu_0}{\sigma_0^2}+\frac{\bar{x}}{\sigma^2/n}\right)\mu\right]} h_1(x_1, \ldots, x_n, \sigma^2, \mu_0, \sigma_0^2)$$

$$f(x_1, x_2, \ldots, x_n, \mu) = e^{-(1/2)\left(\frac{1}{\sigma_0^2}+\frac{1}{\sigma^2/n}\right)\left[\mu-\left(\frac{(\sigma^2/n)\mu_0+\bar{x}\sigma_0^2}{\sigma_0^2+\sigma^2/n}\right)\right]^2} \times h_2(x_1, \ldots, x_n, \sigma^2, \mu_0, \sigma_0^2)$$

$$f(\mu \mid x_1, \ldots, x_n) = e^{-(1/2)\left(\frac{1}{\sigma_0^2}+\frac{1}{\sigma^2/n}\right)\left[\mu-\left(\frac{(\sigma^2/n)\mu_0+\sigma_0^2\bar{x}}{\sigma_0^2+\sigma^2/n}\right)\right]^2} \times h_3(x_1, \ldots, x_n, \sigma^2, \mu_0, \sigma_0^2)$$

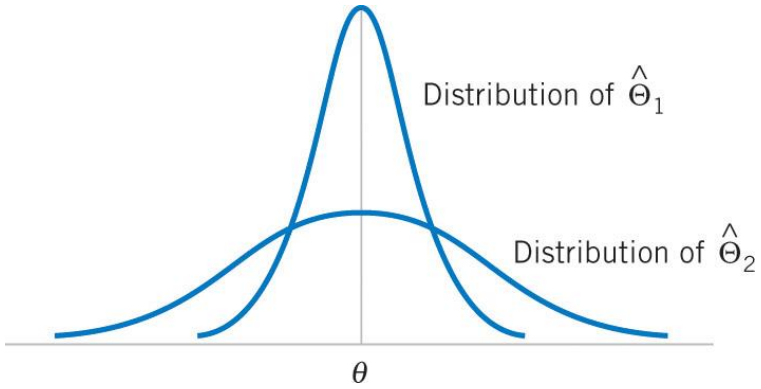
$$\frac{(\sigma^2/n)\mu_0 + \sigma_0^2\bar{x}}{\sigma_0^2 + \sigma^2/n}$$

$$\left(\frac{1}{\sigma_0^2}+\frac{1}{\sigma^2/n}\right)^{-1} = \frac{\sigma_0^2(\sigma^2/n)}{\sigma_0^2+\sigma^2/n}$$

Unbiased estimator

- A statistic $\hat{\theta}$ is said to be an unbiased estimator, or its value an unbiased estimate, *if and only if* the mean of the sampling distribution of the estimator equals θ

$$E(\hat{\theta}) = \theta$$

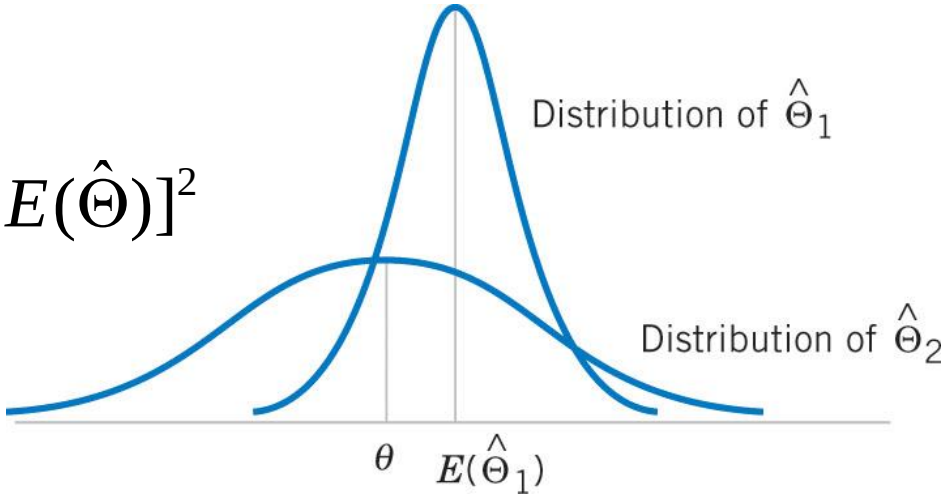


- Exercise
  - Suppose that X is a random variable with mean μ and variance σ^2. Let $X_1, X_2, \ldots X_n$ be a random sample of size *n* from the population represented by X. Show that the sample mean θ and sample variance S^2 are unbiased estimators of μ and σ^2
  - How about S?

Mean squared error均方差

- Minimum variance unbiased estimator
  - If we consider all unbiased estimators of θ, the one with the smallest variance is called the minimum variance un biased estimator(MVUE)
- Mean squared error (MSE) of an estimator

$$MSE(\hat{\Theta}) = E(\hat{\Theta} - \theta)^2$$
$$= E[\hat{\Theta} - E(\hat{\Theta})]^2 + [\theta - E(\hat{\Theta})]^2$$
$$= V(\hat{\Theta}) + (bias)^2$$



A statistics $\hat{\theta}_1$ is said to be a more efficient unbiased estimator of th e parameter than the statistics $\hat{\theta}_2$ if

1. $\hat{\theta}_1$ and $\hat{\theta}_2$ are both unbiased estimators of θ
2. the variance of the sampling distribution of the first estimator is no larger than that of the second and is smaller for at least one v alue of θ.

Some remarks

Under very general and not restrictive conditions, when the sample size n is large and if $\hat{\Theta}$ is the maximum likelihood estimator of the parameter θ,

(1) $\hat{\Theta}$ is an approximately unbiased estimator for θ $[E(\hat{\Theta}) \simeq \theta]$,

(2) the variance of $\hat{\Theta}$ is nearly as small as the variance that could be obtained with any other estimator, and

(3) $\hat{\Theta}$ has an approximate normal distribution.

If $X_1, X_2, \ldots, X_n$ is a random sample of size n from a normal distribution with mean μ and variance σ^2, the sample mean $\overline{X}$ is the MVUE for μ.

Loss function optimality

- Loss function
  - A nonnegative function that generally increases as the di stance between a and θ increases.
  - Two commonly used loss functions
    - Absolute error loss $L(\theta, a) = |a - \theta|$
    - Squared error loss $L(\theta, a) = (a - \theta)^2$
- Risk function
  - The average loss while estimator $\hat{\Theta}$ is used

$$R(\theta, \hat{\Theta}) = E_\theta L(\theta, \hat{\Theta})$$

  - For square error loss

$$R\left(\theta, \widehat{\Theta}\right) = Var_\theta \widehat{\Theta}(\mathbf{X}) + (E_\theta(\widehat{\Theta}) - \theta)^2$$
$$= Var_\theta \widehat{\Theta}(\mathbf{X}) + (Bias_\theta(\widehat{\Theta}))^2$$

Sampling distribution (抽样分布)

- The probability distribution of a statistic is called a sampling distribution
  - For example, let $X_1, \ldots, X_n$ represent n random variables, the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

    - Depends on population distribution, the size of samples and the method of choosing the samples.
- The sampling distribution of $\bar{X}$

Central Limit Theorem: If $\bar{X}$ is the mean of a random sample of size n taken from a population with mean μ and finite variance σ^2, then the limiting form of the distribution of

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}},$$

as $n \to \infty$, is the standard normal distribution $n(z; 0, 1)$.

Four specified distribution

- Z (standard normal), t, chi-square, and F
  - Z and t are closely related to the sampling distribution of means
  - chi-square （χ^2，卡方） and F are closely related to the sampling distribution of variances
- Z distribution

Central Limit Theorem: If $\bar{X}$ is the mean of a random sample of size n taken from a population with mean μ and finite variance σ^2, then the limiting form of the distribution of

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}},$$

as $n \to \infty$, is the standard normal distribution $n(z; 0, 1)$.

  - If n>=30, the approximation for normal will be generally good
  - If n<30, the approximation is good only if the population is not too different from a normal distribution

Sampling Distribution of the Difference between Two Means

If independent samples of size n_1 and n_2 are drawn at random from two populations, discrete or continuous, with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively, then the sampling distribution of the differences of means, $\bar{X}_1 - \bar{X}_2$, is approximately normally distributed with mean and variance given by

$$\mu_{\bar{X}_1-\bar{X}_2} = \mu_1 - \mu_2 \text{ and } \sigma^2_{\bar{X}_1-\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Hence,

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$$

is approximately a standard normal variable.

- Example
  - Two independent experiments are run in which two different types of paint are compared. Eighteen specimens are painted using type *A*, and the drying time, in hours, is recorded for each. The same is done with type *B*. The population standard deviations are both known to be 1.0. Assuming that the mean drying time is equal for the two types of paint, find $P(\bar{x}_A - \bar{x}_B > 1.0)$

Chi-square distribution

$$z = \frac{(X - \mu)}{\sigma}$$

$$z^2 = \chi^2_{(1)}$$

- What if we took 2 values of z^2 at random and added them?

$$z_1^2 = \frac{(x_1 - \mu)^2}{\sigma^2}; z_2^2 = \frac{(x_2 - \mu)^2}{\sigma^2}$$

$$\chi^2_{(2)} = \frac{(x_1 - \mu)^2}{\sigma^2} + \frac{(x_2 - \mu)^2}{\sigma^2} = z_1^2 + z_2^2$$

If $X_1, X_2, \ldots, X_n$ are independent random variables and X_i follows a normal distribution with mean μ_i and variance σ_i^2 for $i = 1, 2, \ldots, n$, then the random variable

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

has a chi-squared distribution with $v = n$ degrees of freedom.

Sample variance: Chi-square distribution

- Sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

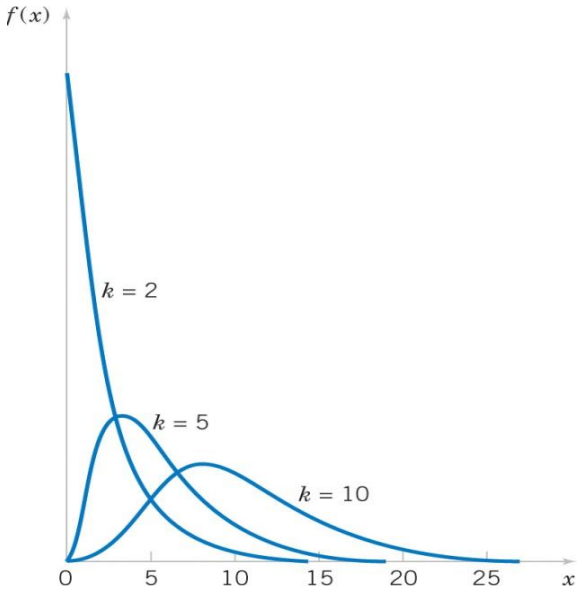
If S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2, then the statistic

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n}\frac{(X_i - \bar{X})^2}{\sigma^2}$$

has a chi-squared distribution with $v = n - 1$ degrees of freedom.

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \mu)^2 &= \sum_{i=1}^{n}[(X_i - \bar{X}) + (\bar{X} - \mu)]^2 \\ &= \sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(\bar{X} - \mu)^2 + 2(\bar{X} - \mu)\sum_{i=1}^{n}(X_i - \bar{X}) \\ &= \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2.\end{aligned}$$

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 = \frac{(n-1)S^2}{\sigma^2} + \frac{(\bar{X} - \mu)^2}{\sigma^2/n}$$

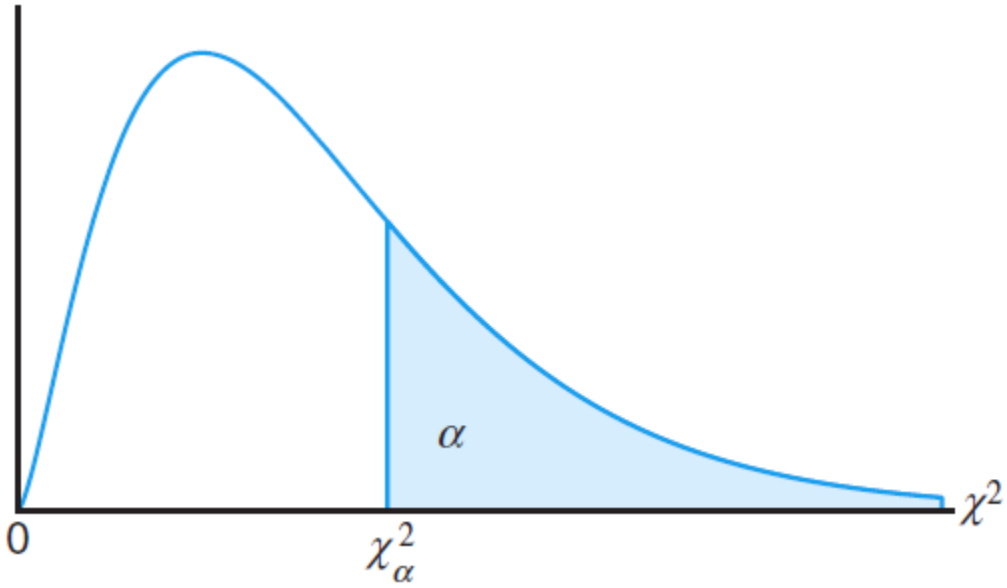


Chi square distribution

- Example
  - A manufacturer of car batteries guarantees that the batteries will last, on average, 3 years with a standard deviation of 1 ye ar. If five of these batteries have lifetimes of 1.9, 2.4, 3.0, 3.5, and 4.2 years, should the manufacturer still be convinced tha t the batteries have a standard deviation of 1 year? Assume th at the battery lifetime follows a normal distribution.

$$s^2 = \frac{(5)(48.26) - (15)^2}{(5)(4)} = 0.815.$$

$$\chi^2 = \frac{(4)(0.815)}{1} = 3.26$$



95% of the χ^2values with 4 degrees of freedom fall between 0.484 and 11.143

t- distribution (学生分布)

- In practice, we don't have more knowledge of σ
- Let $X_1, X_2,..., X_n$ be a random sample from a normal distribution with mean μ and unknown variance σ^2. The random variable

$$T = \frac{\overline{X} - \mu}{S / \sqrt{n}}$$

has a t distribution with n-1 degree of freedom

  - If n>=30, no big difference between t and standard normal

Let Z be a standard normal random variable and V a chi-squared random variable with v degrees of freedom. If Z and V are independent, then the distribution of the random variable T, where

$$T = \frac{Z}{\sqrt{V/v}},$$

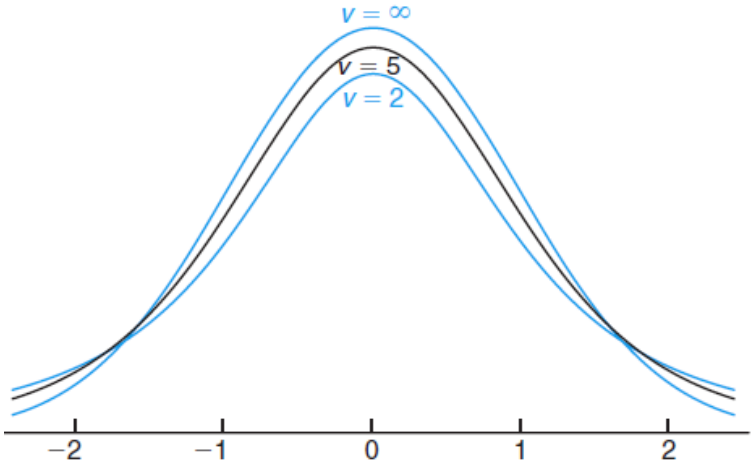
is given by the density function

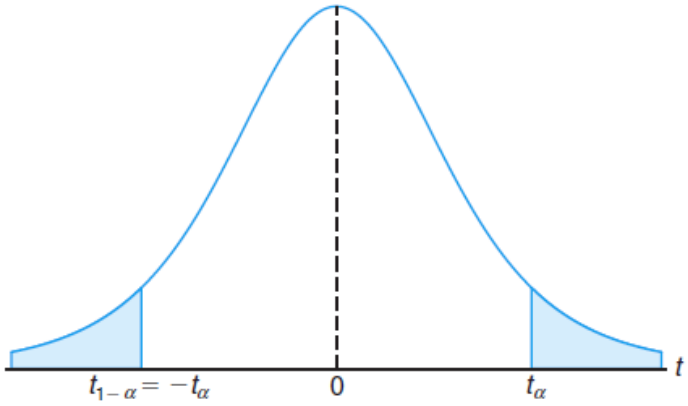
$$h(t) = \frac{\Gamma[(v+1)/2]}{\Gamma(v/2)\sqrt{\pi v}} \left(1 + \frac{t^2}{v}\right)^{-(v+1)/2}, \quad -\infty < t < \infty.$$

This is known as the ***t*-distribution** with v degrees of freedom.

t- distribution

- The *t*-distribution is used extensively in problems that deal with inference about the population mean

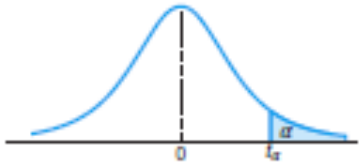




- Example
  - A chemical engineer claims that the population mean yield of a certain batch process is 500 grams per milliliter of raw material. To check this claim he sa mples 25 batches each month. If the computed *t*-value falls between $-t_{0.05}$ and $t_{0.05}$, he is satisfied with this claim. What conclusion should he draw fr om a sample that has a mean $\bar{x}$ = 518 grams per milliliter and a sample stand ard deviation s = 40 grams? Assume the distribution of yields to be approxi mately normal

$$t = \frac{518 - 500}{40/\sqrt{25}} = 2.25,$$

Table A.4 Critical Values of the t-Distribution



	α						
v	**0.40**	**0.30**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179
13	0.259	0.538	0.870	1.079	1.350	1.771	2.160
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064
25	0.256	0.531	0.856	1.058	1.316	1.708	2.060

F-distribution

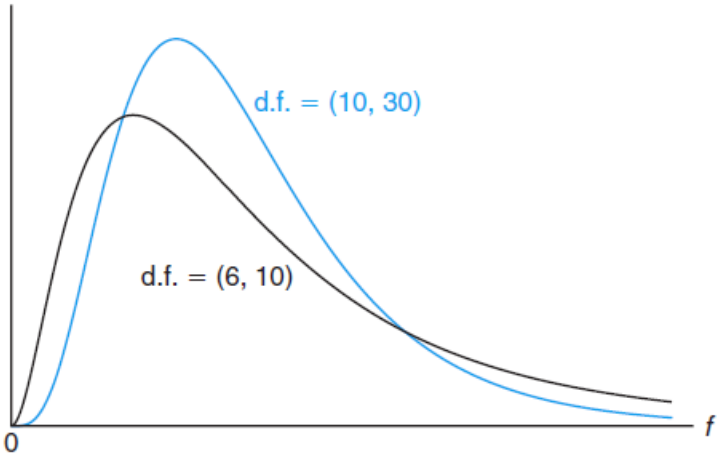
- The statistic F is defined to be the ratio of two independent chi-squared random variables, each divided by its number of degrees of freedom

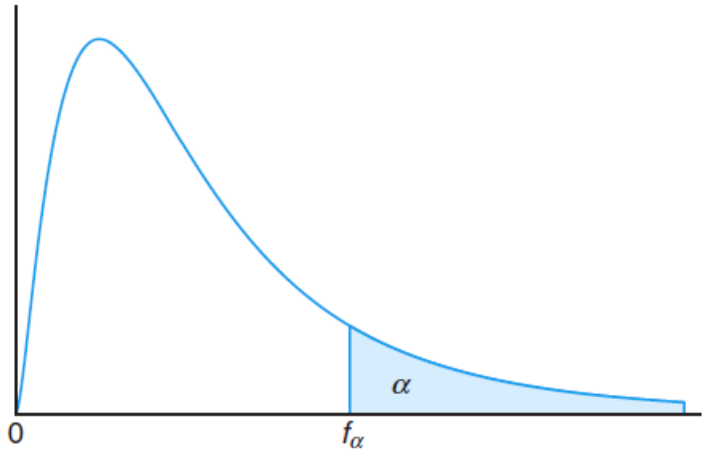
$$F = \frac{U/v_1}{V/v_2}$$

Let U and V be two independent random variables having chi-squared distributions with v_1 and v_2 degrees of freedom, respectively. Then the distribution of the random variable $F = \frac{U/v_1}{V/v_2}$ is given by the density function

$$h(f) = \begin{cases} \frac{\Gamma[(v_1+v_2)/2](v_1/v_2)^{v_1/2}}{\Gamma(v_1/2)\Gamma(v_2/2)} \frac{f^{(v_1/2)-1}}{(1+v_1 f/v_2)^{(v_1+v_2)/2}}, & f > 0, \\ 0, & f \leq 0. \end{cases}$$

This is known as the ***F*-distribution** with v_1 and v_2 degrees of freedom (d.f.).





F-distribution

Writing $f_\alpha(v_1, v_2)$ for f_α with v_1 and v_2 degrees of freedom, we obtain

$$f_{1-\alpha}(v_1, v_2) = \frac{1}{f_\alpha(v_2, v_1)}.$$

If S_1^2 and S_2^2 are the variances of independent random samples of size n_1 and n_2 taken from normal populations with variances σ_1^2 and σ_2^2, respectively, then

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$$

has an F-distribution with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom.

Three things in mind

- One cannot use the Central Limit Theorem unless σ is known. When σ is not known, it should be replaced by s, the sample standard deviation, in order to use the Central Limit Theorem.
- The T statistic is **not** a result of the Central Limit Theorem and $x_1, x_2, \ldots, x_n$ must come from a $n(x; \mu, \sigma)$ distribution in order for $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ to be a t-distribution.
- While the notion of **degrees of freedom** is new at this point, the concept should be very intuitive, since it is reasonable that the nature of the distribution of S and also t should depend on the amount of information in the sample $x_1, x_2, \ldots, x_n$.

Summary

- Descriptive statistics
  - Numerical representations, tables, graphs
  - Scatter plot, histogram, box plot, probability plot
- Point estimation
  - Moment estimation
  - Maximum likelihood estimation
  - Bayesian estimation
- Evaluating the estimators
  - Biasness
- Sampling distributions
  - Z distribution
  - Chi-square distribution
  - T-distribution
  - F-distribution

Statistical Intervals & Hypothesis Tests

Probability and Mathematical Statistics

(概率与数理统计)

Xi ZHANG

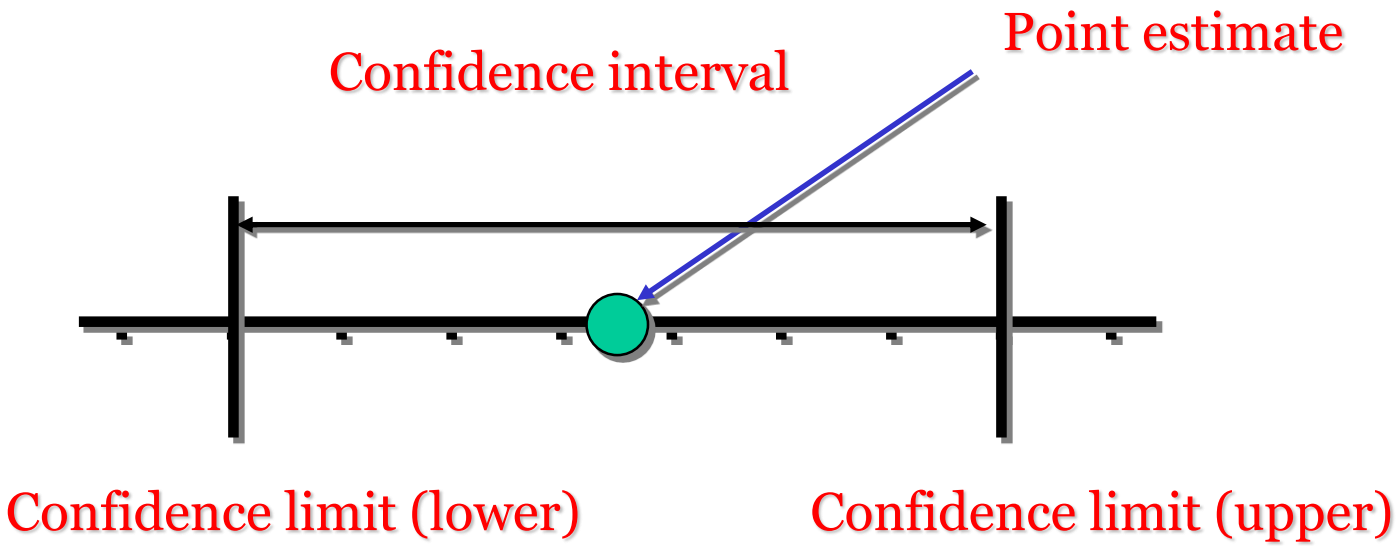
College of Engineering

Outline

- Confidence interval (区间估计)
  - Introduction
  - Confidence interval on mean of normal distribution
  - Confidence interval on variance of normal distribution
  - Confidence interval on population proportion
- Hypotheses testing（假设检验）
  - Introduction
  - Tests on mean of normal distribution
  - Tests on variance of normal distribution
  - Tests on population proportion
  - Contingency table tests
- Statistical inference for two samples（双样本统计推断）
  - Inference on difference in means of two normal distributions
  - Paired t-test
  - Inference on variances of two normal distributions
  - Inference on two population proportions

Interval Estimation

- Provides Range of Values
  - Based on Observations
- Gives Information about Closeness to Unknown Population Parameter
  - Stated in terms of Probability
- Example: Unknown attendance mean with 95% confidence (lies between 50 & 70)



Confidence Interval

- Confidence interval (置信区间，CI) is an interval estimate for a population parameter from samples
- A confidence estimate for μ is an interval with end-points of random variables L and U (置信上限和下限) so that

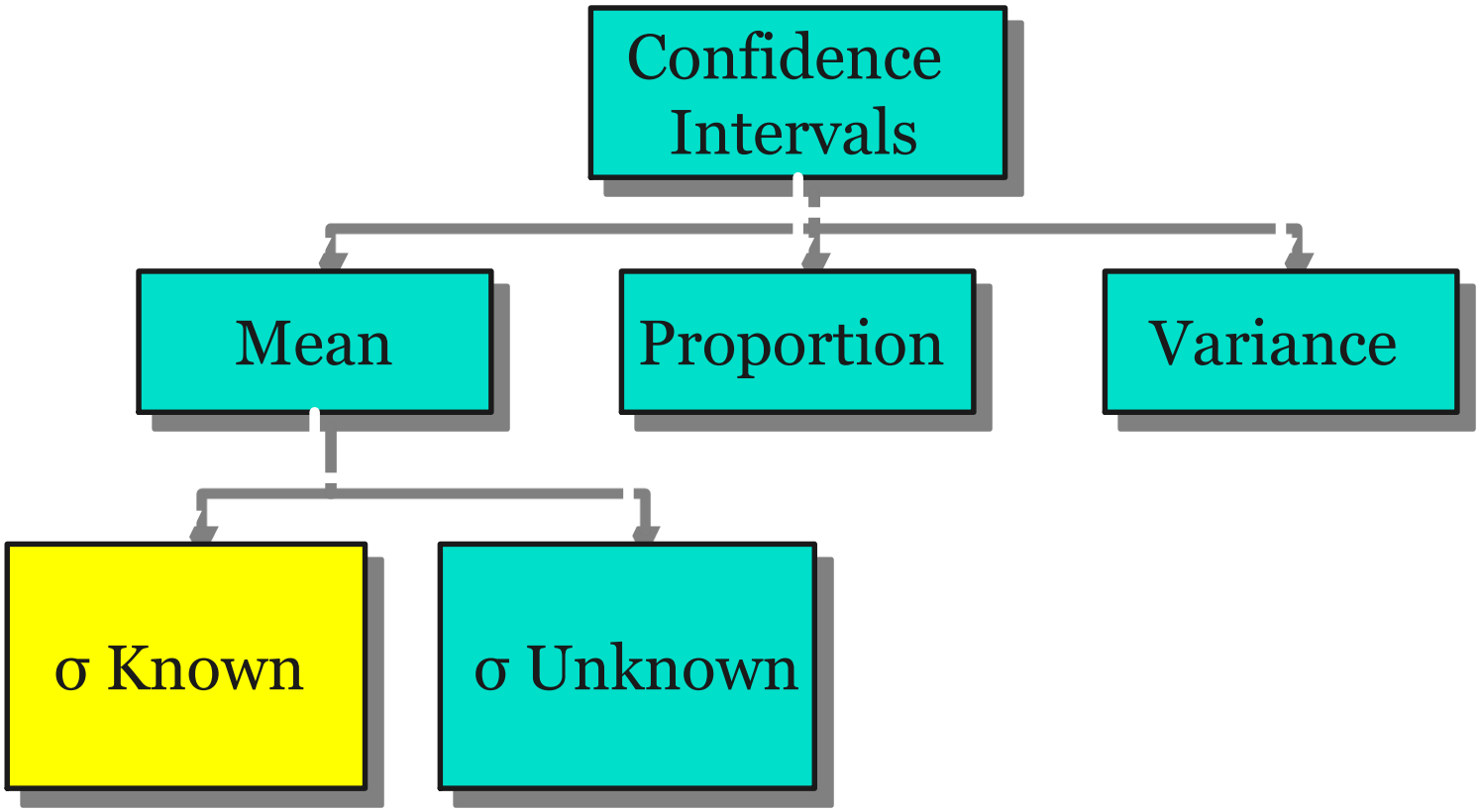
$$P\{L \le \mu \le U\} = 1 - \alpha$$

- Once we have selected the sample, the values of L and U can be estimated, say l and u, then resulting confidence interval for μ is

$$l \le \mu \le u$$

- l and u are lower- and upper-confidence limits; 1-α is called the confidence coefficient (置信系数或置信度)
- The length of a CI is u-l, which is a measure of the precision of estimation

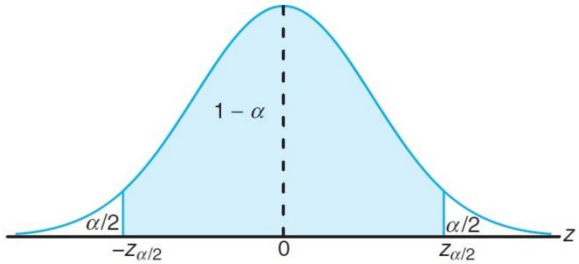
Interval Estimation



CI on Mean of Normal Distribution – Known Variance

- Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a (normal) distribution with unknown mean μ and known variance σ^2, then

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$



has a standard normal distribution (generally $n \geq 30$)

- The probability

$$P\left\{-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq z_{\alpha/2}\right\} = 1 - \alpha$$

A confidence interval can be constructed by

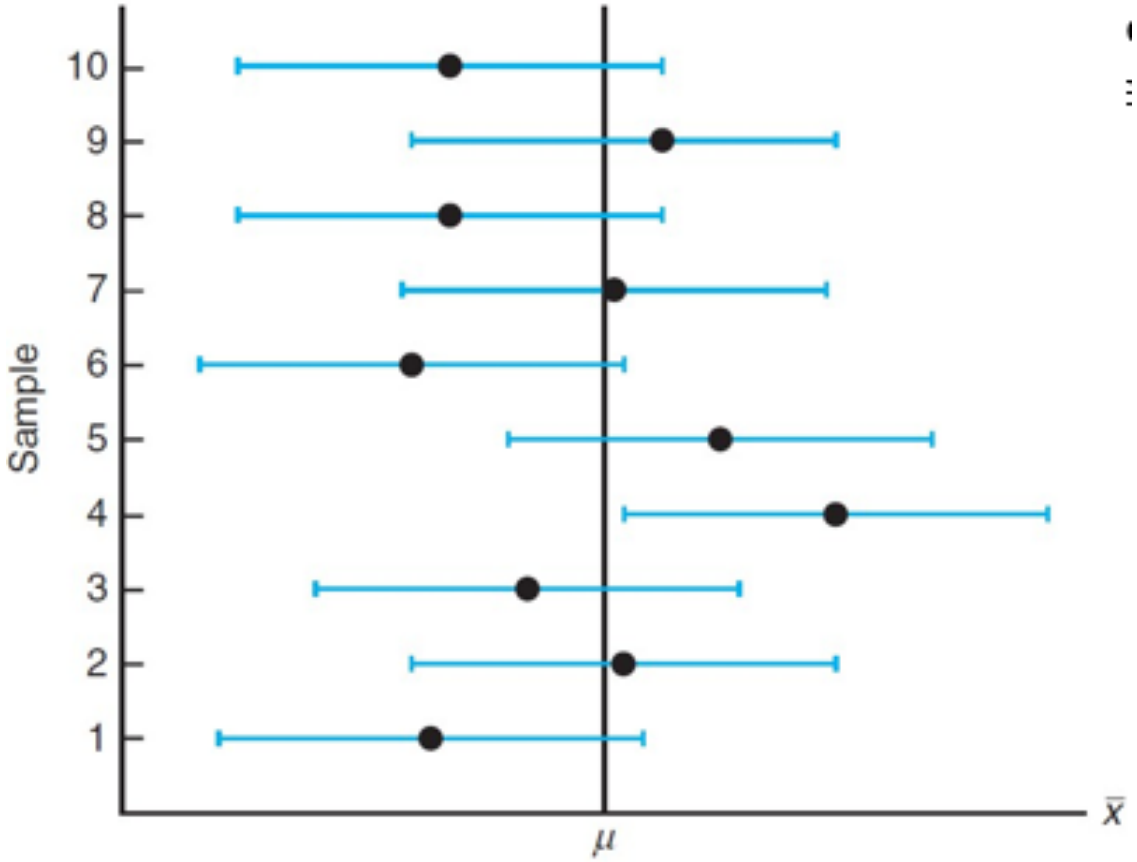
$$P\left\{\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha$$

- Thus a 100(1-α)% CI on μ is given by

$$\bar{x} - z_{\alpha/2}\sigma / \sqrt{n} \leq \mu \leq \bar{x} + z_{\alpha/2}\sigma / \sqrt{n}$$

Confidence Interval

- Interpreting a confidence interval
  - If an infinite number of random samples are collected and a 100(1-α)% confidence interval for μ is computed for each sample, 100(1-α)% of these intervals will contain the true value of μ

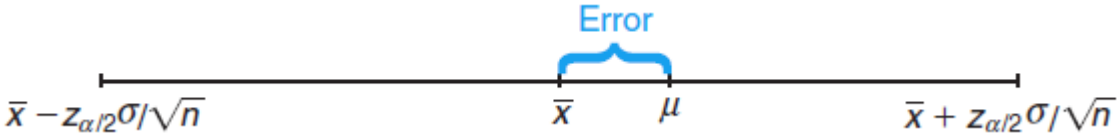


CI on Mean of Normal Distribution – Known Variance

- The precision of the CI is

$$2z_{\alpha/2}\sigma/\sqrt{n}$$

- Error $E = |\bar{x} - \mu|$



- For a specified error E, by setting

$$z_{\alpha/2}\sigma/\sqrt{n} = E$$

we get

$$n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2$$

- Thus if $\bar{x}$ is used as an estimate of μ, we can be 100(1-α) % confident that the error $|\bar{x} - \mu|$ will not exceed E wh en the sample size is set to n

CI on Mean of Normal Distribution – Known Variance

- Two-sided CI

$$\bar{x} - z_{\alpha/2}\sigma/\sqrt{n} \le \mu \le \bar{x} + z_{a/2}\sigma/\sqrt{n}$$

- Upper-confidence bound (one-sided CI, l=-∞)

$$\mu \le \bar{x} + z_{a}\sigma/\sqrt{n}$$

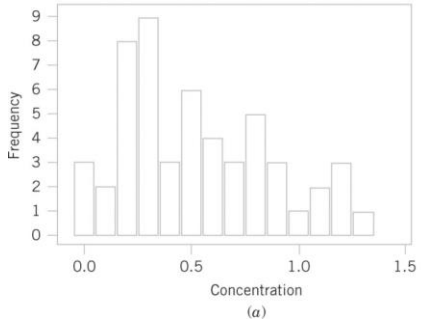
- Lower-confidence bound (one-sided CI, u=∞)

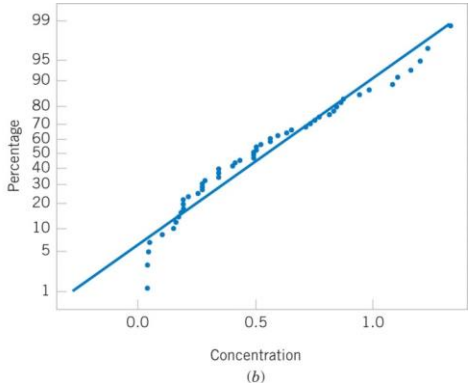
$$\bar{x} - z_{\alpha}\sigma/\sqrt{n} \le \mu$$

CI on Mean of Normal Distribution – Known Variance

- Example
  - A study to investigate the mercury contamination in large mouth bass（大嘴鲈鱼）. A sample of fish was selected from 53 Florida lakes and mercury concentration in the muscle tissue was measur ed, historical data shows the standard deviation of mercury conce ntration is 0.3846. The mercury concentration values are

1.230	0.490	0.490	1.080	0.590	0.280	0.180	0.100	0.940
1.330	0.190	1.160	0.980	0.340	0.340	0.190	0.210	0.400
0.040	0.830	0.050	0.630	0.340	0.750	0.040	0.860	0.430
0.044	0.810	0.150	0.560	0.840	0.870	0.490	0.520	0.250
1.200	0.710	0.190	0.410	0.500	0.560	1.100	0.650	0.270
0.270	0.500	0.770	0.730	0.340	0.170	0.160	0.270	





$$\bar{x} - z_{\alpha/2}\sigma/\sqrt{n} \le \mu \le \bar{x} + z_{a/2}\sigma/\sqrt{n}$$

$$0.5250 - 1.96\frac{0.3486}{\sqrt{53}} \le \mu \le 0.5250 + 1.96\frac{0.3486}{\sqrt{53}}$$

$$0.4311 \le \mu \le 0.6189$$

- Exercise
  - The average zinc (锌) concentration recovered from a sample of measurements taken in 36 different locations in a river is found to be 2.6 grams per milliliter. Find the 95 % and 99% confidence intervals for the mean zinc concentration in the river. Assume that the population standard deviation is 0.3 gram per milliliter.
  - How large a sample is required if we want to be 95% confident that our estimate of μ is off by less than 0.05?

95%

$$2.6 - (1.96)\left(\frac{0.3}{\sqrt{36}}\right) < \mu < 2.6 + (1.96)\left(\frac{0.3}{\sqrt{36}}\right)$$

99%

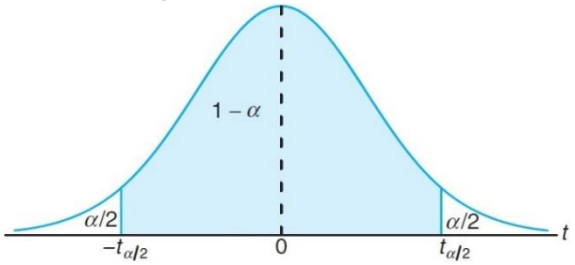
$$2.6 - (2.575)\left(\frac{0.3}{\sqrt{36}}\right) < \mu < 2.6 + (2.575)\left(\frac{0.3}{\sqrt{36}}\right)$$

$$n = \left[\frac{(1.96)(0.3)}{0.05}\right]^2 = 138.3$$

CI on Mean of Normal Distribution – Unknown Variance

- Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a n ormal distribution with unknown mean μ and unknow n variance σ^2, then

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$



has a t distribution with n-1 degrees of freedom

- Since

$$P\left\{-t_{\alpha/2,n-1} \le \frac{\bar{X} - \mu}{S/\sqrt{n}} \le t_{\alpha/2,n-1}\right\} = 1 - \alpha$$

- Then

$$P\left(\bar{X} - t_{\alpha/2,n-1} S/\sqrt{n} \le \mu \le \bar{X} + t_{\alpha/2,n-1} S/\sqrt{n}\right) = 1 - \alpha$$

- Thus if $\bar{x}$ and s are mean and standard deviation, a 100 (1-α)% CI on μ is

$$\bar{x} - t_{\alpha/2,n-1} s/\sqrt{n} \le \mu \le \bar{x} + t_{a/2,n-1} s/\sqrt{n}$$

CI on Mean of Normal Distribution – Unknown Variance

- If the sample size is large ($n \geq 30$), the quantity

$$\frac{\overline{X} - \mu}{S / \sqrt{n}}$$

has an approximate standard normal distribution

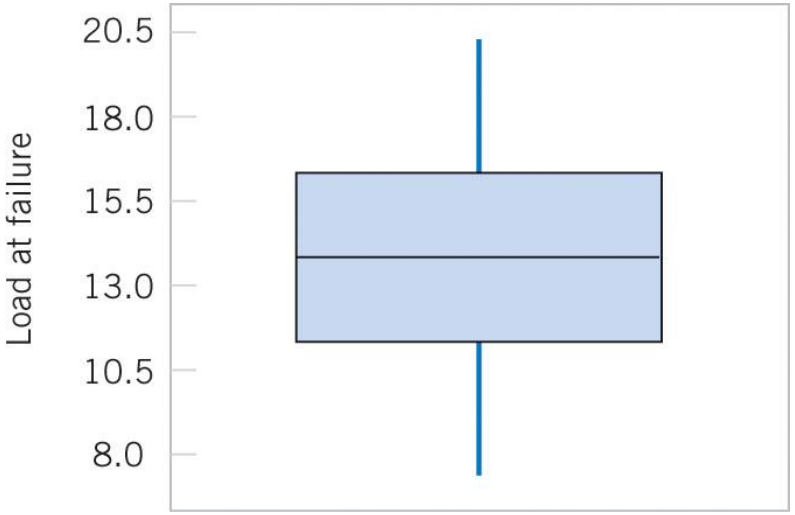
- Thus the following is a large sample CI for μ, with confi dence level of approximately 100(1-α)%

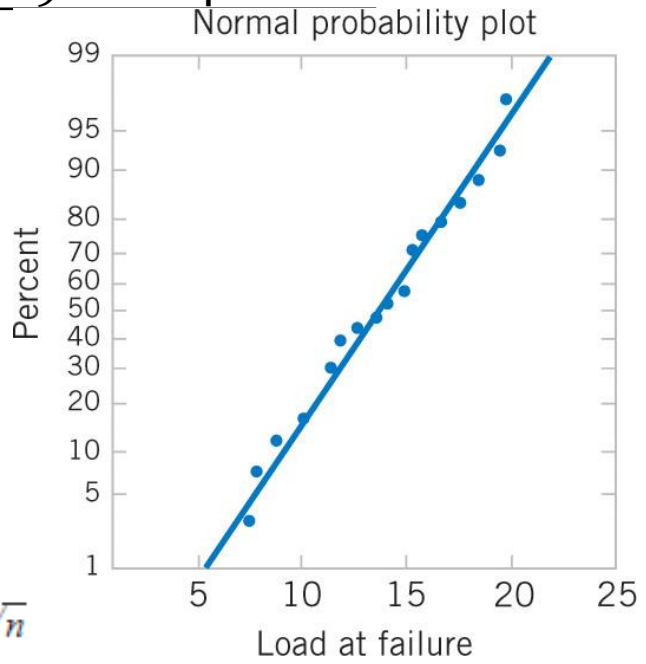
$$\overline{x} - z_{\alpha/2} s / \sqrt{n} \leq \mu \leq \overline{x} + z_{a/2} s / \sqrt{n}$$

CI on Mean of Normal Distribution – Unknown Variance

- Example
  - Tensile adhesion tests on 22 U-700 alloy specimens. The load at specimen failure is as follows

19.8	19.8	14.9	7.5	15.4	15.4
15.4	15.4	7.9	12.7	11.9	11.4
11.4	11.4	17.6	16.7	15.8	
19.5	19.5	13.6	11.9	11.4	





$$\bar{x} - t_{\alpha/2,n-1}s/\sqrt{n} \le \mu \le \bar{x} + t_{\alpha/2,n-1}s/\sqrt{n}$$

$$13.71 - 2.080(3.55)/\sqrt{22} \le \mu \le 13.71 + 2.080(3.55)/\sqrt{22}$$

$$13.71 - 1.57 \le \mu \le 13.71 + 1.57$$

$$12.14 \le \mu \le 15.28$$

Large-Sample Confidence Interval

- Suppose θ is a parameter of a probability distribution, and let $\hat{\Theta}$ be an estimator of θ. If
  - $\hat{\Theta}$ has an approximate normal distribution
  - $\hat{\Theta}$ is approximately unbiased for θ
  - $\hat{\Theta}$ has standard deviation $\sigma_{\hat{\Theta}}$ can be estimated from samp le data
- Then $(\hat{\Theta} - \theta)/\sigma_{\hat{\Theta}}$ is approximately standard normal. A large-sample approximate CI for θ is

$$\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\Theta}} \le \theta \le \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\Theta}}$$

- Maximum likelihood estimators usually satisfy the thre e conditions

CI on Variance of Normal Distribution

- Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a n ormal distribution with mean μ and unknown variance σ^2, and let S^2 be the sample variance, then

$$X^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

has a chi-square (χ^2) distribution with n-1 degrees of f reedom

- Since

$$P\left\{-\chi_{1-\alpha/2,n-1} \le \frac{(n-1)S^2}{\sigma^2} \le \chi_{\alpha/2,n-1}\right\} = 1-\alpha$$

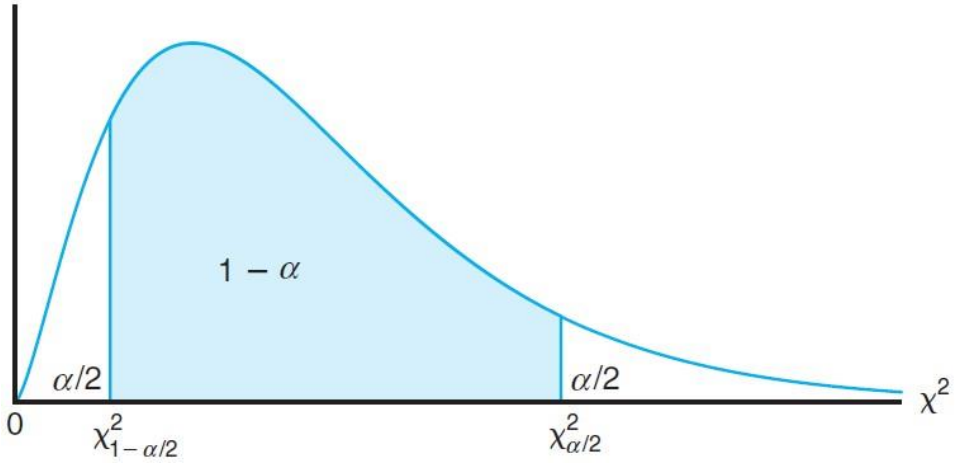
- We get

$$P\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}\right) = 1-\alpha$$

CI on Variance of Normal Distribution

- Two-sided CI on σ^2 is

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}$$



- Furthermore, lower and upper confidence bound on σ^2 is

$$\frac{(n-1)s^2}{\chi^2_{\alpha,n-1}} \le \sigma^2 \quad \text{and} \quad \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha,n-1}}$$

- Example
  - An automatic filling machine is used to fill bottles with li quid detergent (洗涤剂). A random sample of 20 bottles results in a sample variance of fill volume of $s^2 = 0.0153$ $(\text{oz})^2$. If the variance of fill volume is too large, an unacc eptable proportion of bottles will be under- or overfilled. We will assume that the fill volume is approximately nor mally distributed. A 95% upper-confidence interval is

$$\sigma^2 \le \frac{(n-1)s^2}{\chi^2_{0.95,19}} = \frac{(19)0.0153}{10.117} = 0.0287\,(oz)^2$$

CI on Population Proportion

- Suppose that a random sample of size n has been taken from a large population and that X ($\leq$n) observations i n the sample belong to a class of interest. Suppose the proportion of this class in the population is p. Then

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\hat{P} - p}{\sqrt{p(1-p)/n}}$$

  is approximately stand normal

- Confidence interval

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- Choice of sample size

$$n = \left(\frac{z_{\alpha/2}}{E}\right)^2 p(1-p)$$

- Example
  - In a random sample of 85 automobile engine crankshaft bearings, 10 have a surface finish that is rougher than th e specifications allow. Therefore, a point estimate of the proportion of bearings in the population that exceeds the roughness specification is $p = x/n = 10/85 = 0.12$. A 95% two-sided confidence interval for p is

$$\hat{p} - z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$0.12 - 1.96\sqrt{\frac{0.12(0.88)}{85}} \le p \le 0.12 + 1.96\sqrt{\frac{0.12(0.88)}{85}}$$

$$0.05 \le p \le 0.19$$

Hypothesis Testing

- A statistical hypothesis is a statement about the parameters of one or more populations
- Example of hypothesis:

$$H_0 : \mu = 50$$

$$H_1 : \mu \neq 50$$

- H_0: $\mu = 50$ is called the null hypothesis (原假设); H_1: $\mu \neq 50$ is called the alternative hypothesis (备择假设，two-sided)
- One-sided alternative hypothesis

$$H_1 : \mu < 50 \quad \text{or} \quad H_1 : \mu > 50$$

- Hypothesis testing involves taking a random sample, computing a test statistic from the sample data, and make a decision about the null hypothesis

- Suppose that if $48.5 \le \bar{x} \le 51.5$, we will not reject the null hypothesis H_0: $\mu = 50$; otherwise we will reject the null hypothesis. The interval [48.5, 51.5] is called acceptance region; values of $\bar{x}$ less than 48.5 and greater than 51.5 constitute the critical region (临界域). The boundaries between the critical regions and the acceptance region are called the critical value (临界值)
- Type I error （第一类错误）
  - Rejecting the null hypothesis H_0 when it is true
  - Type I error probability is called significance level

$$\alpha = P(\text{type I error}) = P(\text{reject } H_0 \text{ when } H_0 \text{ is true})$$

- Type II error （第二类错误）
  - Failing to reject the null hypothesis when it is false

$$\beta = P(\text{type II error}) = P(\text{fail to reject } H_0 \text{ when } H_0 \text{ is false})$$

	H_0 is true	H_0 is false
Do not reject H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

Hypothesis Testing

- The power (势) of a statistical test is the probability of rejecting the null hypothesis when the alternative hypo thesis is true

$$\text{Power} = 1 - \beta$$

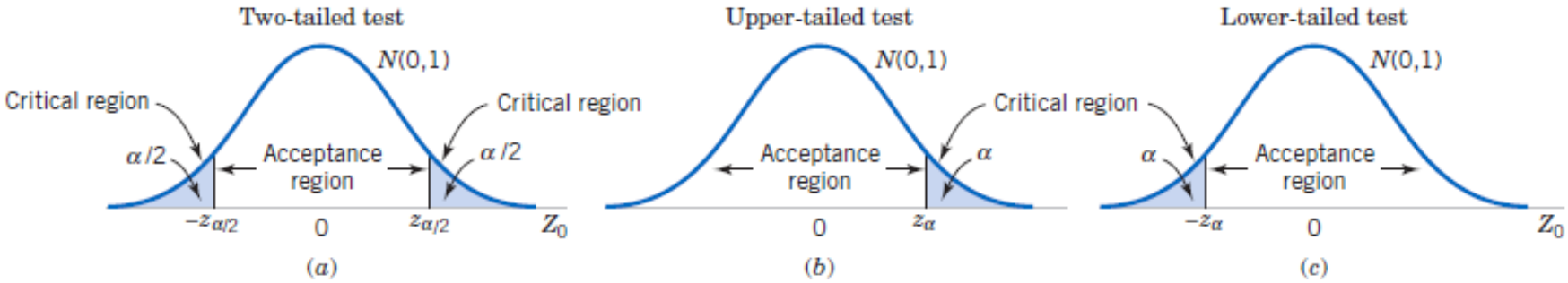
- The *P*-value is the smallest level of significance that wo uld lead to rejection of the null hypothesis with the giv en data
  - The smaller *P*-value is, the stronger of the rejection

1. **Parameter of interest:** From the problem context, identify the parameter of interest.
2. **Null hypothesis, H_0:** State the null hypothesis, H_0.
3. **Alternative hypothesis, H_1:** Specify an appropriate alternative hypothesis, H_1.
4. **Test statistic:** Determine an appropriate test statistic.
5. **Reject H_0 if:** State the rejection criteria for the null hypothesis.
6. **Computations:** Compute any necessary sample quantities, substitute these into the equation for the test statistic, and compute that value.
7. **Draw conclusions:** Decide whether or not H_0 should be rejected and report that in the problem context.

Tests on Mean of Normal Distribution – Known Variance

- Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a n ormal distribution with unknown mean μ and known v ariance σ^2
- We want to test: $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$
- Test statistic: $$Z_0 = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$
- If H_0 is true, then the distribution of Z_0 is $N(0,1)$. Thus the rejection region or critical region is

$$z_0 > z_{\alpha/2} \quad \text{or} \quad z_0 < -z_{\alpha/2}$$



- P-value is $$P = 2\left[1 - \Phi\left(|\, z_0 \,|\right)\right]$$

Tests on Mean of Normal Distribution – Known Variance

- If we want to test

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu > \mu_0$$

Rejection region $z_0 > z_\alpha$

P-value is $P = 1 - \Phi(z_0)$

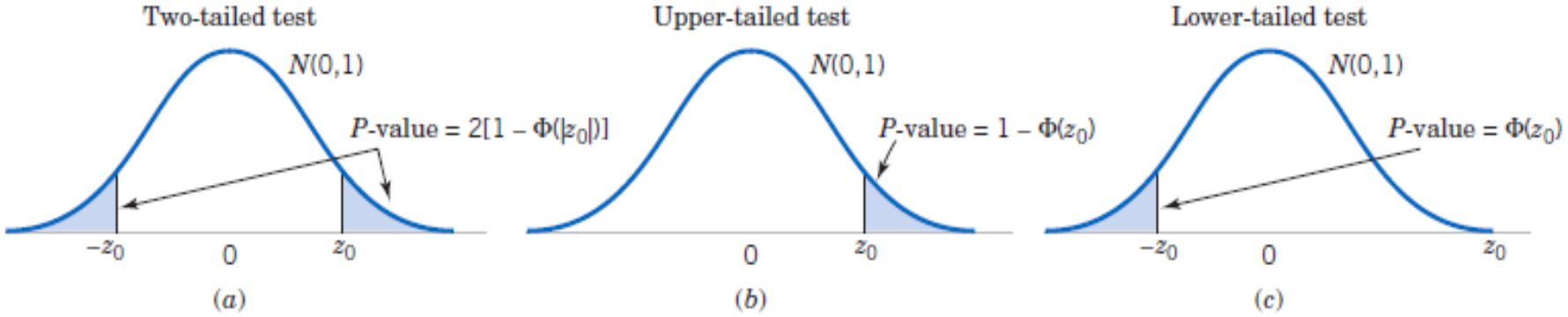
- If we want to test

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu < \mu_0$$

Then rejection region $z_0 < -z_\alpha$

P-value is $P = \Phi(z_0)$



- Consider two-sided hypothesis

$$H_0 : \mu = \mu_0$$

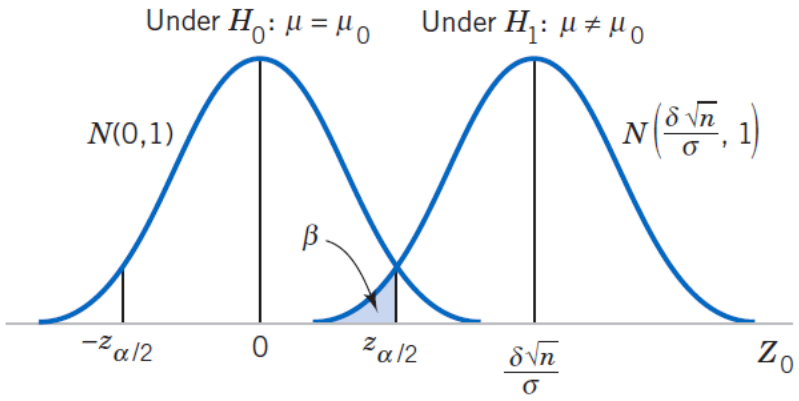
$$H_1 : \mu \neq \mu_0$$

  - Suppose the null hypothesis is false and the true value of the mean is $\mu = \mu_0 + \delta$, where $\delta > 0$, the test statistic

$$Z_0 = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{\bar{X} - (\mu_0 + \delta)}{\sigma / \sqrt{n}} + \frac{\delta\sqrt{n}}{\sigma}$$

  - when H_1 is true

$$Z_0 \sim N\left(\frac{\delta\sqrt{n}}{\sigma}, 1\right)$$



  - Thus

$$\beta = \Phi\left(z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right) - \Phi\left(-z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right)$$

- Since

$$\beta \approx \Phi\left(z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma} \right)$$

- Sample size for two-sided test on the mean, variance known is

$$n \approx \frac{\left(z_{\alpha/2} + z_{\beta}\right)^2 \sigma^2}{\delta^2} \quad \text{where} \quad \delta = \mu - \mu_0$$

- Sample size for one-sided test on the mean, variance known is

$$n \approx \frac{\left(z_{\alpha} + z_{\beta}\right)^2 \sigma^2}{\delta^2} \quad \text{where} \quad \delta = \mu - \mu_0$$

Tests on Mean of Normal Distribution – Known Variance

- Example
  - A random sample of 100 recorded deaths in the United S tates during the past year showed an average life span of 71.8 years. Assuming a population standard deviation of 8.9 years, does this seem to indicate that the mean life sp an today is greater than 70 years? Use a 0.05 level of sig nificance.
- Exercise
  - A manufacturer of sports equipment has developed a ne w synthetic fishing line that the company claims has a m ean breaking strength of 8 kilograms with a standard de viation of 0.5 kilogram. Test the hypothesis that $\mu = 8$ kil ograms against the alternative that $\mu \neq 8$ kilograms if a r andom sample of 50 lines is tested and found to have a mean breaking strength of 7.8 kilograms. Use a 0.01 leve l of significance.

Tests on Mean of Normal Distribution – Unknown Variance

- Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution with unknown mean μ and unknown variance σ^2
- We want to test:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

- Test statistic:

$$T_0 = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$$

- If H_0 is true, then the distribution of T_0 is t distribution with n-1 degrees of freedom. Thus the rejection region or critical region is

$$t_0 > t_{\alpha/2,n-1} \quad \text{or} \quad t_0 < -t_{\alpha/2,n-1}$$

- Rejection region is $t_0 > t_{\alpha,n-1}$ if H_1 is $\mu > \mu_0$; and $t_0 < -t_{\alpha,n-1}$ if H_1 is $\mu < \mu_0$

- Example
  - The increased availability of light materials with high strength has revolutionized the design and manufacture of golf clubs (高球杆), particularly drivers. Clubs with hollow heads and very thin faces can result in much longer tee shots (开球), especially for players of modest skills. This is due partly to the "spring-like effect" that the thin face imparts to the ball. Firing a golf ball at the head of the club and measuring the ratio of the outgoing velocity of the ball to the incoming velocity can quantify this spring-like effect. The ratio of velocities is called the coefficient of restitution of the club. An experiment was performed in which 15 drivers produced by a particular club maker were selected at random and their coefficients of restitution measured. In the experiment the golf balls were fired from an air cannon so that the incoming velocity and spin rate of the ball could be precisely controlled, and we would like to test the mean coefficient of restitution exceeds 0.82. (with $\alpha=0.05$)

0.8411	0.8191	0.8182	0.8125	0.8750
0.8580	0.8532	0.8483	0.8276	0.7983
0.8042	0.8730	0.8282	0.8359	0.8660

1. **Parameter of interest:** The parameter of interest is the mean coefficient of restitution, μ.
2. **Null hypothesis:** H_0: $\mu = 0.82$
3. **Alternative hypothesis:** H_1: $\mu > 0.82$ We want to reject H_0 if the mean coefficient of restitution exceeds 0.82.
4. **Test Statistic:** The test statistic is

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

5. **Reject H_0 if :** Reject H_0 if the P-value is less than 0.05.
6. **Computations:** Since $\bar{x} = 0.83725$, $s = 0.02456$, $\mu_0 = 0.82$, and $n = 15$, we have

$$t_0 = \frac{0.83725 - 0.82}{0.02456/\sqrt{15}} = 2.72$$

$$t_{0.05,14} = 1.761$$

- Exercise
  - The Edison Electric Institute has published figures on the number of kilowatt hours used annually by various home appliances. It is claimed that a vacuum cleaner uses an average of 46 kilowatt hours per year. If a random sample of 12 homes included in a planned study indicates that vacuum cleaners use an average of 42 kilowatt hours per year with a standard deviation of 11.9 kilowatt hours, does this suggest at the 0.05 level of significance that vacuum cleaners use, on average, less than 46 kilowatt hours annually? Assume the population of kilowatt hours to be normal.

1. H_0: $\mu = 46$ kilowatt hours.
2. H_1: $\mu < 46$ kilowatt hours.
3. $\alpha = 0.05$.
4. Critical region: $t < -1.796$, where $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ with 11 degrees of freedom.
5. Computations: $\bar{x} = 42$ kilowatt hours, $s = 11.9$ kilowatt hours, and $n = 12$. Hence,

$$t = \frac{42 - 46}{11.9/\sqrt{12}} = -1.16, \qquad P = P(T < -1.16) \approx 0.135.$$

6. Decision: Do not reject H_0 and conclude that the average number of kilowatt hours used annually by home vacuum cleaners is not significantly less than 46.

Tests on Variance of Normal Distribution

- Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a n ormal distribution with unknown variance σ^2
- We want to test:

$$H_0 : \sigma^2 = \sigma_0^2$$

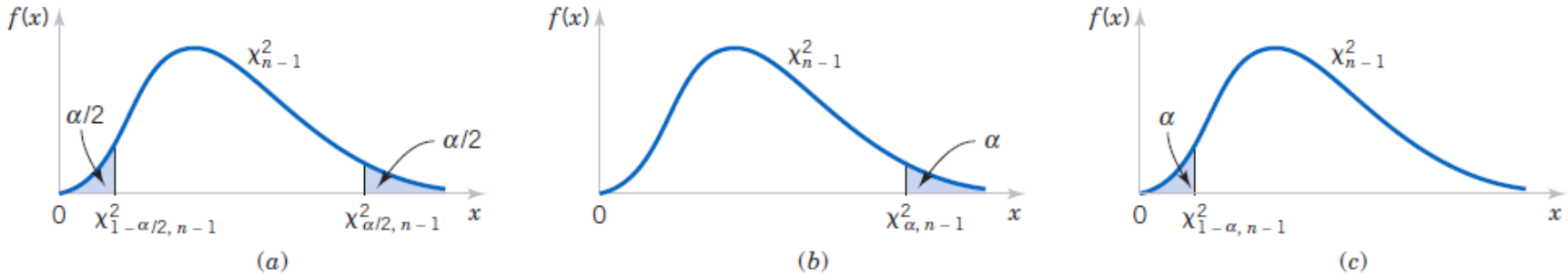
$$H_1 : \sigma^2 \neq \sigma_0^2$$

- Test statistic:

$$\mathrm{X}_0^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

- If H_0 is true, then the distribution of X_0^2 is chi-square d istribution with n-1 degrees of freedom. Thus the reject ion region or critical region is

$$\chi_0^2 > \chi_{\alpha/2,n-1}^2 \quad \text{or} \quad \chi_0^2 < -\chi_{1-\alpha/2,n-1}^2$$



- Example
  - An automatic filling machine is used to fill bottles with liquid d etergent. A random sample of 20 bottles results in a sample var iance of fill volume of $s^2 = 0.0153$ (fluid ounces)2. If the varia nce of fill volume exceeds 0.01 (fluid ounces)², an unacceptable proportion of bottles will be underfilled or overfilled. Is there e vidence in the sample data to suggest that the manufacturer ha s a problem with underfilled or overfilled bottles?

1. **Parameter of Interest:** The parameter of interest is the population variance σ^2.
2. **Null hypothesis:** H_0: $\sigma^2 = 0.01$
3. **Alternative hypothesis:** H_1: $\sigma^2 > 0.01$
4. **Test statistic:** The test statistic is

$$\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

$$\chi_0^2 = 29.07 < \chi_{0.05,19}^2 = 30.14$$

5. **Reject H_0:** Use $\alpha = 0.05$, and reject H_0 if $\chi_0^2 > \chi_{0.05,19}^2 = 30.14$.
6. **Computations:**

$$\chi_0^2 = \frac{19(0.0153)}{0.01} = 29.07$$

Tests on Population Proportion (比例检验)

- For a random variable that follows the binomial distrib ution
- We want to test:

$$H_0 : p = p_0$$

$$H_1 : p \neq p_0$$

- Test statistic （when n is large enough）:

$$Z_0 = \frac{X - np_0}{\sqrt{np_0(1-p_0)}}$$

- If H_0 is true, then the distribution of Z_0 is standard nor mal distribution. Thus the rejection region or critical r egion is

$$z_0 > z_{\alpha/2} \quad \text{or} \quad z_0 < -z_{\alpha/2}$$

Tests on Population Proportion

- Type II error:
  - Suppose p is the true value of the population proportion

$$\beta = \Phi\left(\frac{p_0 - p + z_{\alpha/2}\sqrt{p_0(1-p_0)/n}}{\sqrt{p(1-p)/n}}\right) - \Phi\left(\frac{p_0 - p - z_{\alpha/2}\sqrt{p_0(1-p_0)/n}}{\sqrt{p(1-p)/n}}\right)$$

- Choice of sample size

$$n = \left[\frac{z_{\alpha/2}\sqrt{p_0(1-p_0)} + z_{\beta}\sqrt{p(1-p)}}{p - p_0}\right]^2$$

- Example
  - 某卫视声称全国有70% 的电视观众在周五晚上观看了某吐槽类节目，若某第三方收视机构调查了15家住户观看电视情况，发现其中有8家收看。那么你是否同意这家卫视的这项声明。假设这里我们选用0.10的显著性水平。

1. H_0: $p = 0.7$.
2. H_1: $p \neq 0.7$.
3. $\alpha = 0.10$.
4. Test statistic: Binomial variable X with $p = 0.7$ and $n = 15$.
5. Computations: $x = 8$ and $np_0 = (15)(0.7) = 10.5$.
   the computed P-value is

$$P = 2P(X \leq 8 \text{ when } p = 0.7) = 2\sum_{x=0}^{8} b(x; 15, 0.7) = 0.2622 > 0.10.$$

6. Decision: Do not reject H_0.

Tests on Population Proportion

- Example
  - A semiconductor manufacturer produces controllers used in au tomobile engine applications. The customer requires that the p rocess fallout or fraction defective at a critical manufacturing st ep not exceed 0.05 and that the manufacturer demonstrate pro cess capability at this level of quality using α=0.05. The semico nductor manufacturer takes a random sample of 200 devices a nd finds that four of them are defective. Can the manufacturer demonstrate process capability for the customer?

1. **Parameter of Interest:** The parameter of interest is the process fraction defective p.
2. **Null hypothesis:** $H_0: p = 0.05$
3. **Alternative hypothesis:** $H_1: p < 0.05$
   This formulation of the problem will allow the manufacturer to make a strong claim about process capability if the null hypothesis $H_0: p = 0.05$ is rejected.
4. The test statistic is

$$z_0 = \frac{x - np_0}{\sqrt{np_0(1 - p_0)}}$$

   where $x = 4$, $n = 200$, and $p_0 = 0.05$.
5. **Reject H_0 if:** Reject $H_0: p = 0.05$ if the p-value is less than 0.05.
6. **Computations:** The test statistic is

$$z_0 = \frac{4 - 200(0.05)}{\sqrt{200(0.05)(0.95)}} = -1.95$$

7. **Conclusions:** Since $z_0 = -1.95$, the P-value is $\Phi(-1.95) = 0.0256$, so we reject H_0 and conclude that the process fraction defective p is less than 0.05.

Tests on Goodness of Fit (拟合优度检验)

- We wish to test the hypothesis that a particular distribution will be satisfactory as a population model
- Suppose n observations are arranged in a frequency histogram, having k bins or class intervals. Let O_i be the observed frequency in the ith class interval. And the expected frequency in the ith class interval is E_i
- Test statistic:

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

- If H_0 is true, then the distribution of χ_0^2 is, approximately, chi-square distribution with k-p-1 degrees of freedom. Thus the rejection region or critical region is

$$\chi_0^2 > \chi_{\alpha,k-p-1}^2$$

p is the number of parameters of the hypothesized distribution estimated by sample statistics

the expected frequency is at least larger than 5

- Example
  - The number of defects in p rinted circuit boards is hyp othesized to follow a Poisso n distribution. A random s ample of n = 60 printed bo ards has been collected, an d the following number of defects observed.

Number of Defects	Observed Frequency
0	32
1	15
2	9
3	4

$$p_1 = P(X = 0) = \frac{e^{-0.75}(0.75)^0}{0!} = 0.472$$

$$p_2 = P(X = 1) = \frac{e^{-0.75}(0.75)^1}{1!} = 0.354$$

$$p_3 = P(X = 2) = \frac{e^{-0.75}(0.75)^2}{2!} = 0.133$$

$$p_4 = P(X \geq 3) = 1 - (p_1 + p_2 + p_3) = 0.041$$

Number of Defects	Probability	Expected Frequency
0	0.472	28.32
1	0.354	21.24
2	0.133	7.98
3 (or more)	0.041	2.46

Number of Defects	Observed Frequency	Expected Frequency
0	32	28.32
1	15	21.24
2 (or more)	13	10.44

Contingency Table Tests (列联表)

- We want to know whether the two methods of classific ation are statistically independent
- Contingency table ($r \times c$): each column is a level by on e classification method, and each row is a level by the o ther classification method

		Columns			
		1	2	…	c
Rows	1	O_{11}	O_{12}	…	O_{1c}
	2	O_{21}	O_{22}	…	O_{2c}
	⋮	⋮	⋮	⋮	⋮
	r	O_{r1}	O_{r2}	…	O_{rc}

- Let O_{ij} be the observed frequency for level i of the first classification method and level j on the second classific ation method

- The expected frequency of each cell, if the two methods are independent, is

$$E_{ij} = \frac{1}{n}\sum_{j=1}^{c} O_{ij} \sum_{i=1}^{r} O_{ij}$$

- Then for large n, the statistic

$$\chi_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

has an approximate chi-square distribution with $(r-1)(c-1)$ degrees of freedom if H_0 is true

- Reject region

$$\chi_0^2 > \chi_{\alpha,(r-1)(c-1)}^2$$

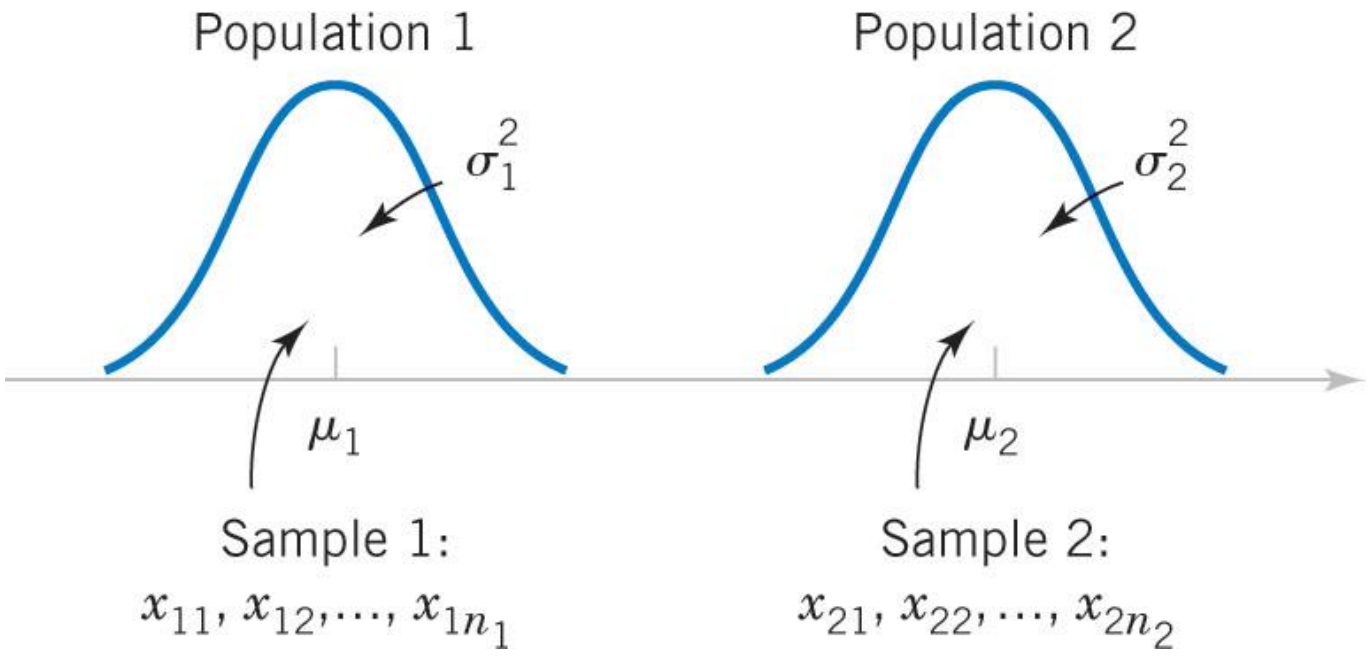
- Example
  - A company has to choose among three pension plans. M anagement wishes to know whether the preference for pl ans is independent of job classification and wants to use α=0.05. The opinions of a random sample of 500 emplo yees are shown as follows:

	Health Insurance Plan			
Job Classification	1	2	3	Totals
Salaried workers	160	140	40	340
Hourly workers	40	60	60	160
Totals	200	200	100	500

	Health Insurance Plan			
Job Classification	1	2	3	Totals
Salaried workers	136	136	68	340
Hourly workers	64	64	32	160
Totals	200	200	100	500

Statistical Inference for Two Samples

- Compare two different population based on two sampl es



Difference in Means of Two Normals – Known Variances

- Suppose that $X_{11}, X_{12}, \ldots, X_{1n_1}$ is a random sample from population 1; and $X_{21}, X_{22}, \ldots, X_{2n_2}$ is a random sample f rom population 2. The two populations are independe nt and $X_1 \sim N(\mu_1, \sigma_1^2), X_2 \sim N(\mu_2, \sigma_2^2)$.
  - Suppose variances are known
  - Then

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Difference in Means of Two Normals – Known Variances

- Hypothesis tests on the difference in means, variances known
  - Null hypothesis: $\mu_1 - \mu_2 = \Delta_0$
  - Test statistic:

$$Z_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

  - Alternative hypothesis

Alternative hypothesis	Rejection region
$H_1 : \mu_1 - \mu_2 \neq \Delta_0$	$z_0 > z_{\alpha/2}$ or $z_0 < z_{-\alpha/2}$
$H_1 : \mu_1 - \mu_2 > \Delta_0$	$z_0 > z_{\alpha}$
$H_1 : \mu_1 - \mu_2 < \Delta_0$	$z_0 < z_{-\alpha}$

Difference in Means of Two Normals – Known Variances

- Hypothesis tests on the difference in means, variances known
  - Type II error

$$\beta = \Phi(z_{\alpha/2} - \frac{\Delta - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}) - \Phi(-z_{\alpha/2} - \frac{\Delta - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}})$$

  - Sample size under equivalence condition ($n_1 = n_2 = n$)

$$n \approx \frac{(z_{\alpha/2} + z_\beta)^2(\sigma_1^2 + \sigma_2^2)}{(\Delta - \Delta_0)^2}$$

Difference in Means of Two Normals – Known Variances

- CI on the difference in means, variances knows
  - Two-sided

$$\bar{x}_1 - \bar{x}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le \bar{x}_1 - \bar{x}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

  - One-sided upper

$$\mu_1 - \mu_2 \le \bar{x}_1 - \bar{x}_2 + z_{\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

  - One-sided lower

$$\bar{x}_1 - \bar{x}_2 - z_{\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2$$

  - Sample size under equivalence condition ($n_1 = n_2 = n$)

$$n = \left(\frac{z_{\alpha/2}}{E}\right)^2 \left(\sigma_1^2 + \sigma_2^2\right)$$

Difference in Means of Two Normals – Known Variances

- Example
  - A product developer is interested in reducing the drying time of a primer paint. Two formulations of the paint are tested; formulation 1 is the standard chemistry, and for mulation 2 has a new drying ingredient that should redu ce the drying time. From experience, it is known that the standard deviation of drying time is 8 minutes, and this i nherent variability should be unaffected by the addition of the new ingredient. 10 specimens are painted with for mulation 1, and another 10 specimens are painted with f ormulation 2; the 20 specimens are painted in random o rder. The two sample average drying times are 121 minut es and 112 minutes, respectively. What conclusions can t he product developer draw about the effectiveness of the new ingredient, using α=0.05?

1. **Parameter of interest:** The quantity of interest is the difference in mean drying times, $\mu_1 - \mu_2$, and $\Delta_0 = 0$.
2. **Non hypothesis:** H_0: $\mu_1 - \mu_2 = 0$, or H_0: $\mu_1 = \mu_2$.
3. **Alternative hypothesis:** H_1: $\mu_1 > \mu_2$. We want to reject H_0 if the new ingredient reduces mean drying time.
4. **Test statistic:** The test statistic is

$$z_0 = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

where $\sigma_1^2 = \sigma_2^2 = (8)^2 = 64$ and $n_1 = n_2 = 10$.

5. **Reject H_0 if:** Reject H_0: $\mu_1 = \mu_2$ if the P-value is less than 0.05.
6. **Computations:** Since $\bar{x}_1 = 121$ minutes and $\bar{x}_2 = 112$ minutes, the test statistic is

$$z_0 = \frac{121 - 112}{\sqrt{\frac{(8)^2}{10} + \frac{(8)^2}{10}}} = 2.52$$

Difference in Means of Two Normals – Unknown Variances

- Suppose that $X_{11}, X_{12}, \ldots, X_{1n_1}$ is a random sample from population 1; and $X_{21}, X_{22}, \ldots, X_{2n_2}$ is a random sample f rom population 2. The two populations are independe nt and $X_1 \sim N(\mu_1, \sigma_1^2), X_2 \sim N(\mu_2, \sigma_2^2)$.
  - Suppose variances are unknown
- Case 1: equal variance $\sigma_1^2 = \sigma_2^2 = \sigma^2$
  - Null hypothesis: $\mu_1 - \mu_2 = \Delta_0$
  - Pooled estimator (合并估计) for σ^2:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

  - Test statistic:

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Difference in Means of Two Normals – Unknown Variances

- Case 1: equal variance $\sigma_1^2 = \sigma_2^2 = \sigma^2$
  - Null hypothesis: $\mu_1 - \mu_2 = \Delta_0$
  - Alternative hypothesis — Rejection region

$$H_1 : \mu_1 - \mu_2 \neq \Delta_0 \qquad t_0 > t_{\alpha/2,n_1+n_2-2} \text{ or } t_0 < t_{\alpha/2,n_1+n_2-2}$$

$$H_1 : \mu_1 - \mu_2 > \Delta_0 \qquad t_0 > t_{\alpha,n_1+n_2-2}$$

$$H_1 : \mu_1 - \mu_2 < \Delta_0 \qquad t_0 < -t_{\alpha,n_1+n_2-2}$$

  - $1-\alpha$ CI for $\mu_1 - \mu_2$

$$\bar{x}_1 - \bar{x}_2 - t_{\alpha/2,n_1+n_2-2} s_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \le \mu_1 - \mu_2 \le \bar{x}_1 - \bar{x}_2 + t_{\alpha/2,n_1+n_2-2} s_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

Difference in Means of Two Normals – Unknown Variances

- Case 2: unequal variance $\sigma_1^2 \neq \sigma_2^2$
  - Test statistic

$$T_0^* = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

  - Degree of freedom

$$\upsilon = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

  - $1 - \alpha$ CI for $\mu_1 - \mu_2$

$$\bar{x}_1 - \bar{x}_2 - t_{\alpha/2,\upsilon}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + t_{\alpha/2,\upsilon}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Difference in Means of Two Normals – Unknown Variances

- Example
  - Two catalysts are being analyzed to determine how they affect the mean yield of a chemical process. Specifically, catalyst 1 is currently in use, but catalyst 2 is acceptable. Since catalyst 2 is cheaper, it should be adopted, providi ng it does not change the process yield. A test is run in th e pilot plant and results in the data shown as below. Is th ere any difference between the mean yields? (Minitab)

Observation Number	Catalyst 1	Catalyst 2
1	91.50	89.19
2	94.18	90.95
3	92.18	90.46
4	95.39	93.21
5	91.79	97.19
6	89.07	97.04
7	94.72	91.07
8	89.21	92.75
	$\bar{x}_1 = 92.255$	$\bar{x}_2 = 92.733$
	$s_1 = 2.39$	$s_2 = 2.98$

- Exercise
  - Arsenic (砷) concentration in public drinking water supp lies is a potential health risk. An article in the *Arizona R epublic* (Sunday, May 27, 2001) reported drinking water arsenic concentrations in parts per billion (ppb) for 10 m ethropolitan Phoenix communities and 10 communities i n rural Arizona. The data follow:

Metro Phoenix ($\bar{x}_1 = 12.5, s_1 = 7.63$)	Rural Arizona ($\bar{x}_2 = 27.5, s_2 = 15.3$)
Phoenix, 3	Rimrock, 48
Chandler, 7	Goodyear, 44
Gilbert, 25	New River, 40
Glendale, 10	Apachie Junction, 38
Mesa, 15	Buckeye, 33
Paradise Valley, 6	Nogales, 21
Peoria, 12	Black Canyon City, 20
Scottsdale, 25	Sedona, 12
Tempe, 15	Payson, 1
Sun City, 7	Casa Grande, 18

Paired *t*-test

- When the observations on the two populations of inter est are collected in pairs.
- Suppose that $(X_{11}, X_{21}), (X_{12}, X_{22}), \ldots, (X_{1n}, X_{2n})$ is a set of n paired observations. Suppose the difference $D_j = X_{1j} - X_{2j}$ are normally distributed with mean

$$\mu_D = \mu_1 - \mu_2$$

- Null hypothesis: $\mu_D = \Delta_0$
- Test statistic:

$$T_0 = \frac{\overline{D} - \Delta_0}{S_D / \sqrt{n}}$$

- Alternative hypothesis — Rejection region

$H_1 : \mu_D \neq \Delta_0$ — $t_0 > t_{\alpha/2,n-1}$ or $t_0 < t_{\alpha/2,n-1}$

$H_1 : \mu_D > \Delta_0$ — $t_0 > t_{\alpha,n-1}$

$H_1 : \mu_D < \Delta_0$ — $t_0 < -t_{\alpha,n-1}$

- Example
  - An article in the *Journal of Strain Analysis* (1983, Vol. 1 8, No. 2) compares several methods for predicting the sh ear strength for steel plate girders （梁）. Data for two o f these methods, the Karlsruhe and Lehigh procedures, w hen applied to nine specific girders, are shown in the Ta ble. We wish to determine whether there is any differenc e (on the average) between the two methods.

Girder	Karlsruhe Method	Lehigh Method	Difference d_j
S1/1	1.186	1.061	0.125
S2/1	1.151	0.992	0.159
S3/1	1.322	1.063	0.259
S4/1	1.339	1.062	0.277
S5/1	1.200	1.065	0.135
S2/1	1.402	1.178	0.224
S2/2	1.365	1.037	0.328
S2/3	1.537	1.086	0.451
S2/4	1.559	1.052	0.507

Inference on Variances of Two Normal Distributions

- Suppose that $X_{11}, X_{12}, \ldots, X_{1n_1}$ is a random sample from population 1; and $X_{21}, X_{22}, \ldots, X_{2n_2}$ is a random sample from population 2. The two populations are independent normal with unknown variances, then

$$F = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} \sim f\left(n_1 - 1, n_2 - 1\right)$$

- Null hypothesis: $\sigma_1^2 = \sigma_2^2$
- Test statistic: $F_0 = S_1^2 / S_2^2$
- Alternative hypothesis — Rejection region

Alternative hypothesis	Rejection region
$H_1 : \sigma_1^2 \neq \sigma_2^2$	$f_0 > f_{\alpha/2, n_1-1, n_2-1}$ or $f_0 < f_{1-\alpha/2, n_1-1, n_2-1}$
$H_1 : \sigma_1^2 > \sigma_2^2$	$f_0 > f_{\alpha, n_1-1, n_2-1}$
$H_1 : \sigma_1^2 < \sigma_2^2$	$f_0 < f_{1-\alpha, n_1-1, n_2-1}$

- CI on the ratio of two variances

$$\frac{s_1^2}{s_2^2} f_{1-\alpha/2, n_2-1, n_1-1} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{s_1^2}{s_2^2} f_{\alpha/2, n_2-1, n_1-1}$$

Inference on Variances of Two Normal Distributions

- Example
  - Oxide layers on semiconductor wafers are etched in a mixture of gases to achieve the proper thickness. The variability in the t hickness of these oxide layers is a critical characteristic of the w afer, and low variability is desirable for subsequent processing steps. Two different mixtures of gases are being studied to dete rmine whether one is superior in reducing the variability of the oxide thickness. Sixteen wafers are etched in each gas. The sam ple standard deviations of oxide thickness are $s_1 = 1.96$ angstr oms and $s_2 = 2.13$ angstroms, respectively. Is there any eviden ce to indicate that either gas is preferable?

$$s_1^2/s_2^2=3.8316/4.5369=0.8445$$

Large-sample Test on Difference in Population Proportions

- Suppose two independent random samples of size n_1 a nd n_2 are taken from two populations, and let X_1, X_2 re present the number of observations that belong to the class of interest in samples 1 and 2, respectively.

$$\hat{P}_1 = X_1 / n_1, \quad \hat{P}_2 = X_2 / n_2$$

- Null hypothesis: $p_1 = p_2$
- Test statistic:

$$Z_0 = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}(1-\hat{P})(\frac{1}{n_1} + \frac{1}{n_2})}} \qquad \hat{P} = \frac{X_1 + X_2}{n_1 + n_2}$$

- Alternative hypothesis — Rejection region

Alternative hypothesis	Rejection region
$H_1 : p_1 \neq p_2$	$z_0 > z_{\alpha/2}$ or $z_0 < z_{-\alpha/2}$
$H_1 : p_1 > p_2$	$z_0 > z_{\alpha}$
$H_1 : p_1 < p_2$	$z_0 < z_{-\alpha}$

Large-sample Test on Difference in Population Proportions

- Type II error
  - When H_0 is false, the standard deviation of $\hat{P}_1 - \hat{P}_2$

$$\sigma_{\hat{P}_1-\hat{P}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

  - Type II error for $H_1 : p_1 \neq p_2$

$$\beta = \Phi\left[\frac{z_{\alpha/2}\sqrt{\bar{p}\bar{q}(1/n_1+1/n_2)} - (p_1-p_2)}{\sigma_{\hat{P}_1-\hat{P}_2}}\right]$$

$$-\Phi\left[\frac{-z_{\alpha/2}\sqrt{\bar{p}\bar{q}(1/n_1+1/n_2)} - (p_1-p_2)}{\sigma_{\hat{P}_1-\hat{P}_2}}\right]$$

While

$$\bar{p} = \frac{n_1p_1 + n_2p_2}{n_1+n_2}$$

$$\bar{q} = \frac{n_1(1-p_1) + n_2(1-p_2)}{n_1+n_2}$$

Large-sample Test on Difference in Population Proportions

- Type II error
  - Type II error for $H_1 : p_1 > p_2$

$$\beta = \Phi\left[\frac{z_\alpha\sqrt{\bar{p}\bar{q}(1/n_1 + 1/n_2)} - (p_1 - p_2)}{\sigma_{\hat{P}_1 - \hat{P}_2}}\right]$$

$$\bar{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \quad \text{and} \quad \bar{q} = \frac{n_1(1 - p_1) + n_2(1 - p_2)}{n_1 + n_2}$$

  - Type II error for $H_1 : p_1 < p_2$

$$\beta = 1 - \Phi\left[\frac{-z_\alpha\sqrt{\bar{p}\bar{q}(1/n_1 + 1/n_2)} - (p_1 - p_2)}{\sigma_{\hat{P}_1 - \hat{P}_2}}\right]$$

- Sample size

$$n = \frac{\left[z_{\alpha/2}\sqrt{(p_1 + p_2)(q_1 + q_2)/2} + z_\beta\sqrt{p_1 q_1 + p_2 q_2}\right]^2}{(p_1 - p_2)^2}$$

- $1 - \alpha$ CI

$$\hat{p}_1 - \hat{p}_2 - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

$$\leq p_1 - p_2 \leq \hat{p}_1 - \hat{p}_2 + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

Large-sample Test on Difference in Population Proportions

- Example
  - Extracts of St. John's Wort（圣约翰草）are widely used to treat depression. An article in the April 18, 2001 issue of *the Journal of the American Medical Association* ("Effectiveness of St. John's Wort on Major Depression: A Randomized Controlled Trial") compared the efficacy of a standard extract of St. John's Wort with a placebo (安慰剂) in 200 outpatients (门诊) diagnosed with major depression. Patients were randomly assigned to two groups; one group received the St. John's Wort, and the other received the placebo. After eight weeks, 19 of the placebo-treated patients showed improvement, whereas 27 of those treated with St. John's Wort improved. Is there any reason to believe that St. John's Wort is effective in treating major depression?

$$Z_0 = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}(1-\hat{P})(\frac{1}{n_1}+\frac{1}{n_2})}} = \frac{0.19\text{-}0.27}{\sqrt{0.23(0.77)\quad(0.01+0.01)}} = -1.34$$

Summary

- Confidence interval
  - Confidence interval on mean of normal distribution

$$\bar{x} - z_{\alpha/2}\sigma/\sqrt{n} \le \mu \le \bar{x} + z_{a/2}\sigma/\sqrt{n}$$

$$\bar{x} - t_{\alpha/2,n-1}s/\sqrt{n} \le \mu \le \bar{x} + t_{a/2,n-1}s/\sqrt{n}$$

  - Confidence interval on variance of normal distribution

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}$$

  - Confidence interval on population proportion

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Summary

- Hypotheses testing
  - Introduction
    - Type I error, type II error, power, *P*-value
  - Tests on mean of normal distribution

$$Z_0 = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \qquad T_0 = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$$

  - Tests on variance of normal distribution

$$\mathrm{X}_0^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

  - Tests on population proportion

$$Z_0 = \frac{X - np_0}{\sqrt{np_0(1-p_0)}}$$

  - Tests on goodness of fit

$$\mathrm{X}_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Summary

- Hypotheses testing
  - Contingency table tests

$$\chi_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

- Statistical inference for two samples
  - Inference on difference in means of two normal distributions
    - Known variance

$$\bar{x}_1 - \bar{x}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le \bar{x}_1 - \bar{x}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

$$Z_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

Summary

- Statistical inference for two samples
  - Inference on difference in means of two normal distributions
    - Unknown variance – case 1: equal variance

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} \qquad T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

$$\bar{x}_1 - \bar{x}_2 - t_{\alpha/2,n_1+n_2-2}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \le \mu_1-\mu_2 \le \bar{x}_1 - \bar{x}_2 + t_{\alpha/2,n_1+n_2-2}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

    - Unknown variance – case 2: unequal variance

$$T_0^* = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \qquad \upsilon = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1}+\frac{(s_2^2/n_2)^2}{n_2-1}}$$

$$\bar{x}_1 - \bar{x}_2 - t_{\alpha/2,\upsilon}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \le \mu_1-\mu_2 \le \bar{x}_1 - \bar{x}_2 + t_{\alpha/2,\upsilon}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

Summary

- Statistical inference for two samples
  - Paired t-test

$$T_0 = \frac{\overline{D} - \Delta_0}{S_D / \sqrt{n}}$$

  - Inference on variances of two normal distributions

$$F_0 = S_1^2 / S_2^2$$

$$\frac{s_1^2}{s_2^2} f_{1-\alpha/2, n_2-1, n_1-1} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2} f_{\alpha/2, n_2-1, n_1-1}$$

  - Inference on two population proportions

$$Z_0 = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}(1-\hat{P})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

$$\hat{p}_1 - \hat{p}_2 - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$\le p_1 - p_2 \le \hat{p}_1 - \hat{p}_2 + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

HW

- Due on Dec. 5th

About course project

- Target: To apply what you have learned in class or off class to solve a problem you are interested
- No specific problem, you can name your own
- Encourage to use statistical software, e.g. R, SAS, Python…
- Team: 2 HW groups form into a single team by Dec. 5th
- Grading:
  - Lightening talk (less than 5+1 mins)+ project report (submit on final exam day, the mainbody should be less than 5 pages)
  - Presentation time: Dec. 22nd and 26th in class
- Signature project will have additional bonus up to 5 credit s (One or two groups entitled)
- Sample exam showing

Sample project topics

- Class Tardiness Between Genders
  - Is there a relationship between gender and tardiness among the students at PK U? On average are male students tardy more often than female students and fo r a longer time period?
- Credit Hours Taken
  - Do junior and seniors take, on average, more credit hours per semester than fr eshmen and sophomores?
- Gender Fuel/Gas Buying Behavior
  - Is there a relationship between gender and the amount of gas pumped during a visit to the gas station?
- Computer Lab Capacity and Peak Time Analyses
  - Is there an issue with the capacity of the engineering computer labs? What is th e impact on the need for student to own a lap top?
- Casual Parking Lot versus Parking garage Usage
  - Which time slot is the busiest for parking? Morning? Afternoon? Evening? Is t here a difference in parking availability among the different lots?
- Library Usage
  - Is there a difference in the usage of library by student level at PKU?
- Popularity of Two Music Bands
  - The comparison of the popularity of two music groups (i.e. Phoenix Legend vs. Mayday). For one band, is there a difference in popularity due to gender or age?

Linear Regression

Probability and Mathematical Statistics

(概率与数理统计)

Xi ZHANG

College of Engineering

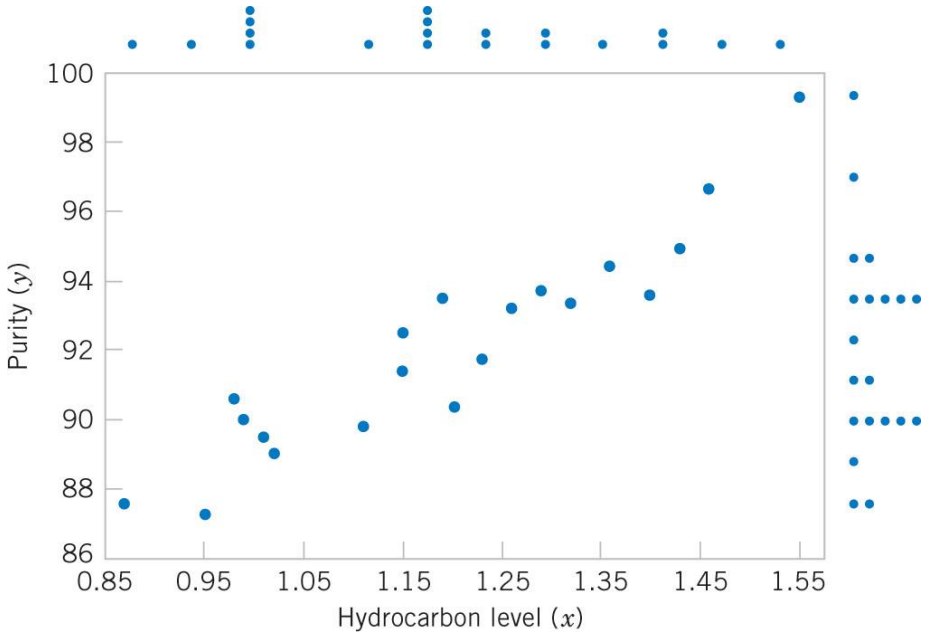
Outline

- Introduction
- Simple linear regression
  - Parameter estimation
  - Hypothesis tests on parameters
  - Confidence intervals on parameters
  - Prediction of new observations
  - Model adequacy
- Multiple linear regression
  - Parameter estimation
  - Hypothesis tests on parameters
  - Confidence intervals
  - Polynomial regression model
  - Categorical regression
  - Variable selection
  - Stepwise regression

Introduction

- Exploring the relationships between two or more varia bles
  - Example: chemical distillation process (化工蒸馏过程)

Observation Number	Hydrocarbon Level x (%)	Purity y (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33



Introduction

- Question: how to learn from data and build empirical models for prediction?

- We have different ‘screwdrivers’
  - In statistics: regression, classification, clustering
  - In engineering: signal processing, pattern recognition
  - In computer science: machine learning, artificial intelligence

Introduction

- Empirical model
  - When internal mechanism is not clear
  - When any simple first principle model to describe compl ex system behavior
- Why empirical model
  - Data could be easier to obtain than years before
  - Computer techs develop rapidly
  - More reliable and confident on data

- Regression model is one of canonical empirical models

Introduction

- General model: $Y = f_s(X)$
  - Predictors:
    - $X = (X_1, X_2, ..., X_p)$
  - Output:
    - Y
- Training data:
  - Consist of a finite of N i.i.d. samples from $p(x, y)$:

$$S = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$$

- Algorithm: to search over a space (hypothesis space), to sele ct f_s

Introduction

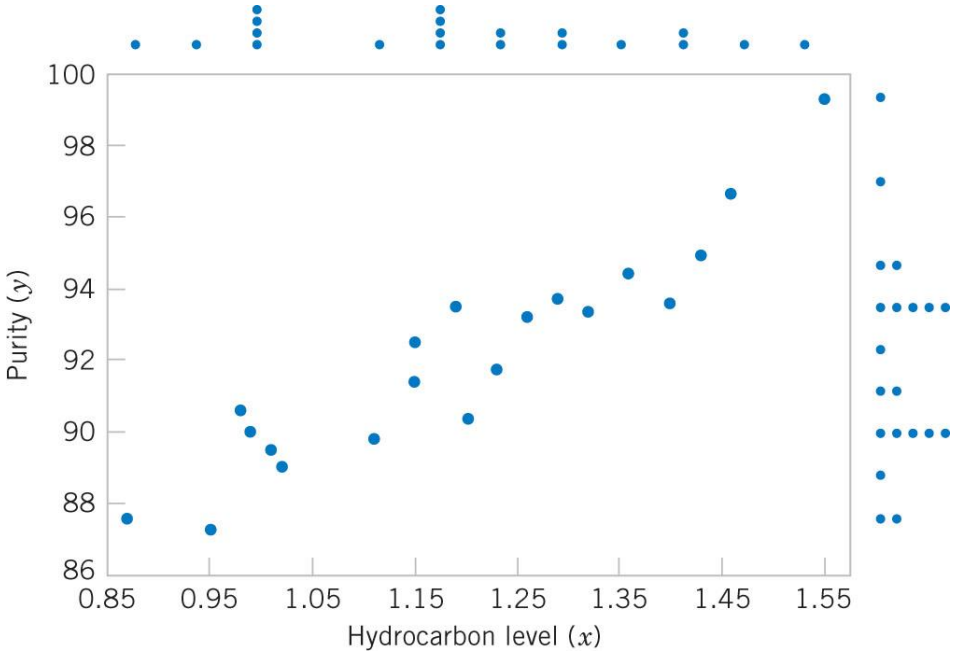
- Linear regression
  - Assume that the expected value of Y is a linear function o f x

$$y = \beta_0 + \beta_1 x + \varepsilon$$

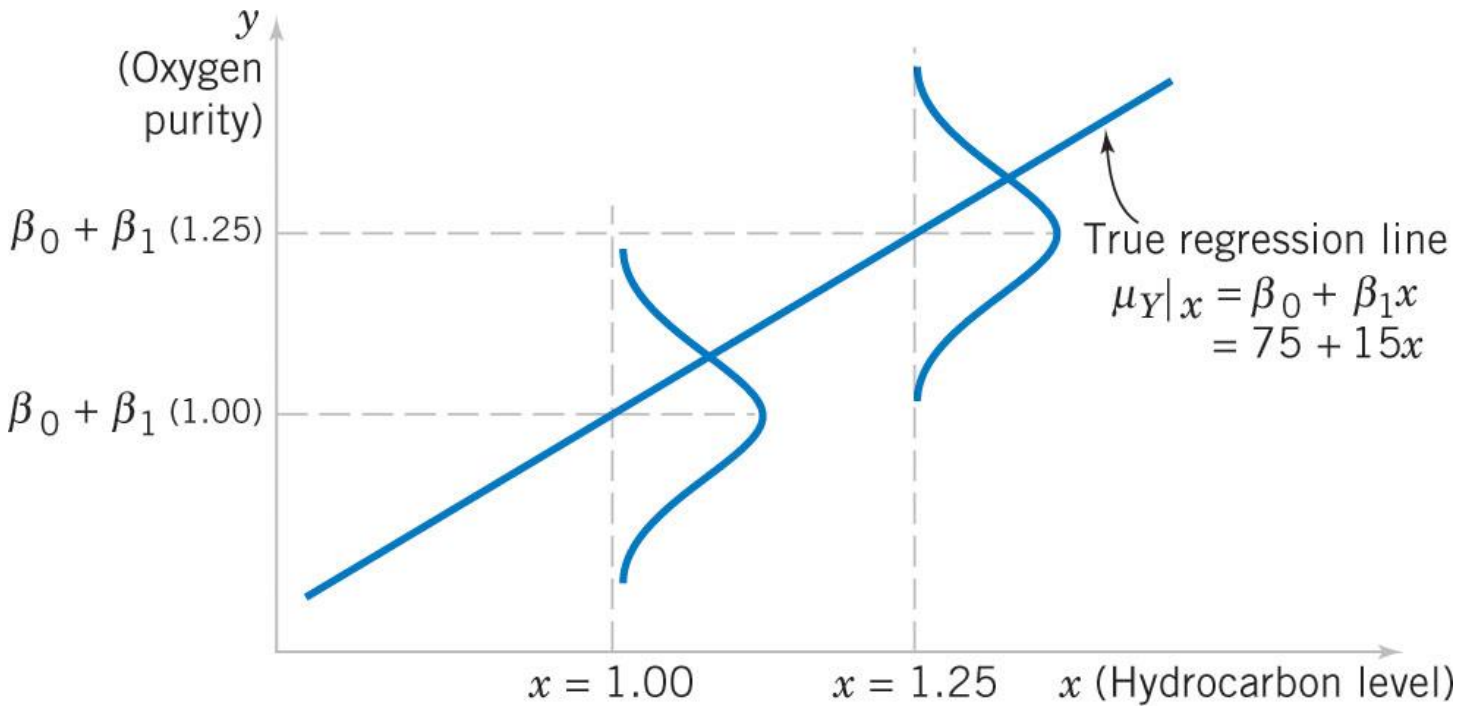
  - Suppose that the mean and variance of ε are 0 and σ^2

$$E(y \mid x) = E(\beta_0 + \beta_1 x + \varepsilon)$$
$$= \beta_0 + \beta_1 x + E(\varepsilon) = \beta_0 + \beta_1 x$$

$$V(y \mid x) = V(\beta_0 + \beta_1 x) + V(\varepsilon)$$
$$= 0 + \sigma^2 = \sigma^2$$



Introduction



$$E(Y|x) = \mu_{Y|x} = \beta_0 + \beta_1 x$$

Simple Linear Regression

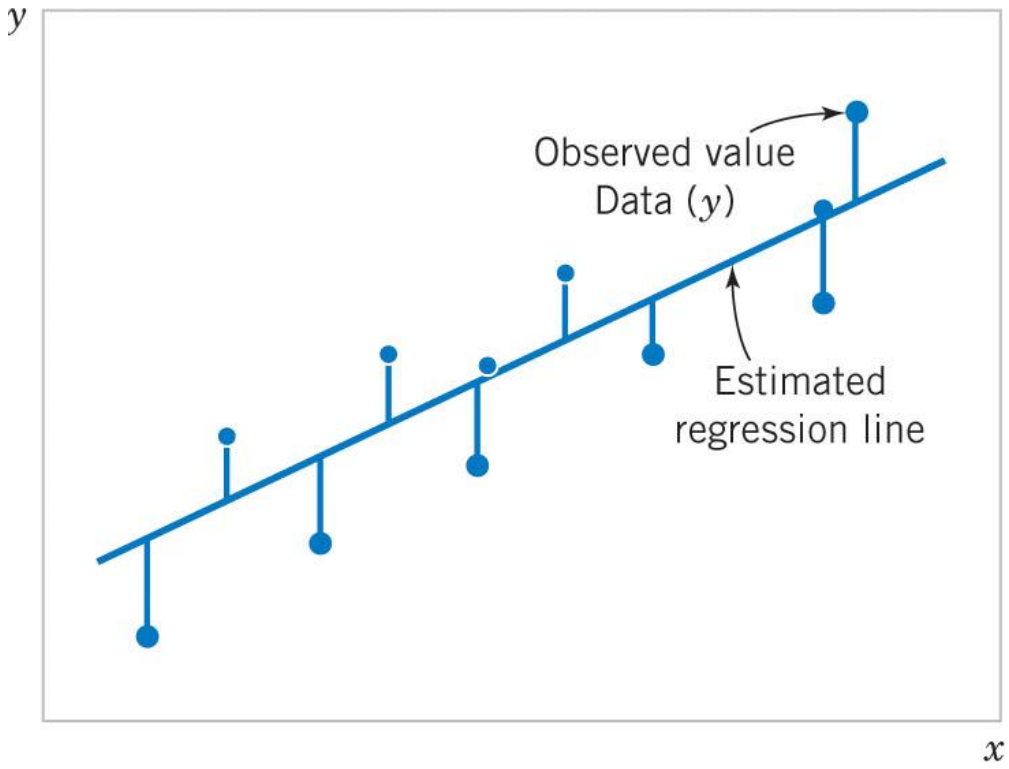
- A single predictor x and a response variable Y

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

- Least squares （最小二乘法）
  - **Minimize** the sum of the squares of the vertical deviation

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon,$$
$$i = 1,2,\ldots, n$$

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$



Simple Linear Regression

- To minimize L, the least squares estimator （LSE）mu st satisfy:

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0,\hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\left.\frac{\partial L}{\partial \beta_1}\right|_{\hat{\beta}_0,\hat{\beta}_1} = -2\hat{\beta}_1\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

- The LSE in the simple linear regression

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}, \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

S_{xy}

S_{xx}

Simple Linear Regression

- Example: chemical distillation process (化工蒸馏过程)

Observation Number	Hydrocarbon Level x (%)	Purity y (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

$$n = 20 \quad \sum_{i=1}^{20} x_i = 23.92 \quad \sum_{i=1}^{20} y_i = 1{,}843.21$$

$$\bar{x} = 1.1960 \quad \bar{y} = 92.1605$$

$$\sum_{i=1}^{20} y_i^2 = 170{,}044.5321 \quad \sum_{i=1}^{20} x_i^2 = 29.2892$$

$$\sum_{i=1}^{20} x_i y_i = 2{,}214.6566$$

$$S_{xx} = \sum_{i=1}^{20} x_i^2 - \frac{\left(\sum_{i=1}^{20} x_i\right)^2}{20} = 29.2892 - \frac{(23.92)^2}{20}$$

$$= 0.68088$$

and

$$S_{xy} = \sum_{i=1}^{20} x_i y_i - \frac{\left(\sum_{i=1}^{20} x_i\right)\left(\sum_{i=1}^{20} y_i\right)}{20}$$

$$= 2{,}214.6566 - \frac{(23.92)(1{,}843.21)}{20} = 10.17744$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{10.17744}{0.68088} = 14.94748$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 92.1605 - (14.94748)1.196 = 74.28331$$

- Solution to Example

Regression Analysis

The regression equation is

Purity = 74.3 + 14.9 HC Level

Predictor	Coef	SE Coef	T	P
Constant	74.283 ← $\hat{\beta}_0$	1.593	46.62	0.000
HC Level	14.947 ← $\hat{\beta}_1$	1.317	11.35	0.000

S = 1.087 R-Sq = 87.7% R-Sq (adj) = 87.1%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	152.13	152.13	128.86	0.000
Residual Error	18	21.25 ← SS_E	1.18 ← $\hat{\sigma}^2$		
Total	19	173.38			

Predicted Values for New Observations

New Obs	Fit	SE Fit	95.0% CI	95.0% PI
1	89.231	0.354	(88.486, 89.975)	(86.830, 91.632)

Values of Predictors for New Observations

New Obs	HC Level
1	1.00

- Estimation of the variance of the error term ε
  - Sum error sum of squares

$$SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

  - Unbiased estimator of variance

$$\hat{\sigma}^2 = \frac{SS_E}{n-2}$$

  - More comfortable way

$$SS_E = SS_T - \hat{\beta}_1 S_{xy}$$

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2 - \hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x}_i)(y_i - \bar{y}_i)$$

Properties of the LSE

- For β_1

$$E(\hat{\beta}_1) = \beta_1$$

$$V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$$

- For β_0

$$E(\hat{\beta}_0) = \beta_0$$

$$V(\hat{\beta}_0) = \sigma^2[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}]$$

$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}])$$

Hypothesis Tests in Simple Linear Regression

- To test the hypothesis that the slope equals to a consta nt

$$H_0 : \beta_1 = \beta_{1,0}$$

$$H_1 : \beta_1 \neq \beta_{1,0}$$

- Test statistic

$$T_0 = \frac{\hat{\beta}_1 - \hat{\beta}_{1,0}}{\sqrt{\hat{\sigma}^2 / S_{xx}}}$$

Hypothesis Tests in Simple Linear Regression

- To test the hypothesis that the intercept equals to a constant

$$H_0 : \beta_0 = \beta_{0,0}$$

$$H_1 : \beta_0 \neq \beta_{0,0}$$

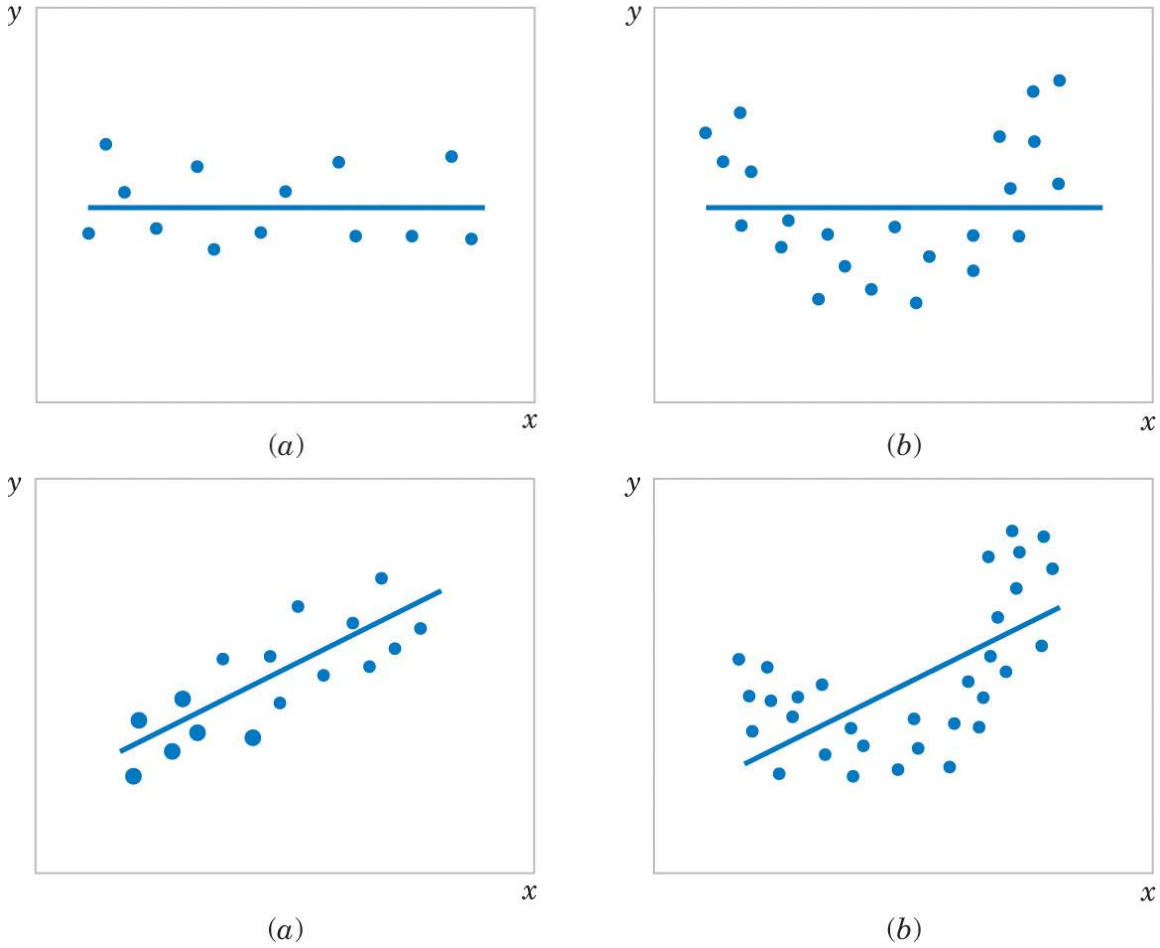
- Test statistic

$$T_0 = \frac{\hat{\beta}_0 - \hat{\beta}_{0,0}}{\sqrt{\hat{\sigma}^2 [\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}]}}$$

Hypothesis Tests in Simple Linear Regression

- A special case: significance of regression

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$



- Example: Chemical distillation process
  - Suppose we want to test for significance of regression model i n previous oxygen purity data

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

  - Our test statistic

$$T_0 = \frac{\hat{\beta}_1 - \hat{\beta}_{1,0}}{\sqrt{\hat{\sigma}^2 / S_{xx}}} = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / S_{xx}}} = \frac{14.947}{\sqrt{1.18 / 0.68088}} = 11.35$$

Confidence Interval

- If the error term ε_i, in the regression model are i.i.d.:

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \qquad \text{and} \qquad \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]}}$$

  are distributed as *t* random variables with *n*-2 DOFs

- CI for β_1

$$\hat{\beta}_1 - t_{\alpha/2,n-2} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2,n-2} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

- CI for β_0

$$\hat{\beta}_0 - t_{\alpha/2,n-2} \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]} \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2,n-2} \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]}$$

Confidence Interval

- Example: Chemical distillation process
  - We will find a 95% CI on the slope of the regression line usin g the data in the previous oxygen purity example

$$\hat{\beta}_1 - t_{0.025,18}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \le \beta_1 \le \hat{\beta}_1 + t_{0.025,18}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

$$14.947 - 2.101\sqrt{\frac{1.18}{0.68088}} \le \beta_1 \le 14.947 + 2.101\sqrt{\frac{1.18}{0.68088}}$$

This simplifies to

$$12.181 \le \beta_1 \le 17.713$$

Confidence Interval

- A point estimate of the mean of Y

$$\hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

- The variance of the estimate is

$$V(\hat{\mu}_{Y|x_0}) = \sigma^2[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}]$$

$$\frac{\hat{\mu}_{Y|x_0} - \mu_{Y|x_0}}{\sqrt{\hat{\sigma}^2[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}]}} \sim t_{n-2}$$

$$\hat{\mu}_{Y|x_0} - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}]} \le \mu_{Y|x_0} \le \hat{\mu}_{Y|x_0} + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}]}$$

Confidence Interval

- Example: Chemical distillation process
  - In the Chemical distillation process example, 95% CI on $\mu_{Y|x_0}$

$$\hat{\mu}_{Y|x_0} \pm 2.1011.18\sqrt{\hat{\sigma}^2[\frac{1}{20}+\frac{(x_0-1.196)^2}{0.68088}]}$$

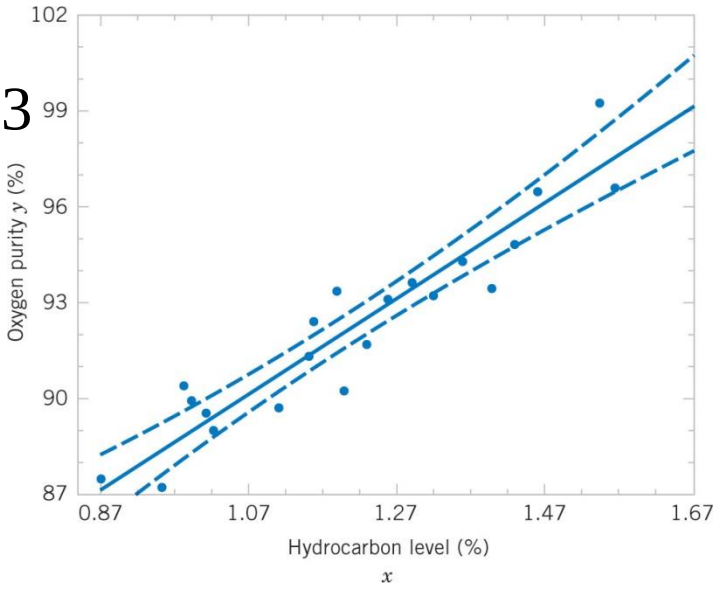
  - When $x_0 = 1\%$

$$\hat{\mu}_{Y|x_{1.00\%}} = 74.283 + 14.947(1.00) = 89.23$$

  - The 95% CI

$$\{89.23 \pm 2.101\sqrt{1.18[\frac{1}{20}+\frac{(1.00-1.1960)^2}{0.68088}]}\}$$

$$88.48 \le \mu_{Y|1.00} \le 89.98$$



Prediction of New Observations

- If x_0 is the value of the regressor variable of interest

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

- The error in prediction $e_{\hat{p}} = Y_0 - \hat{Y}_0$

  - Mean: 0
  - Variance: $V(e_{\hat{p}}) = V(Y_0 - \hat{Y}_0) = \sigma^2[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}]$
  - CI: $$\frac{Y_0 - \hat{Y}_0}{\sqrt{\hat{\sigma}^2[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}]}} \sim t_{n-2}$$

$$\hat{y}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}]} \le Y_0 \le \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}]}$$

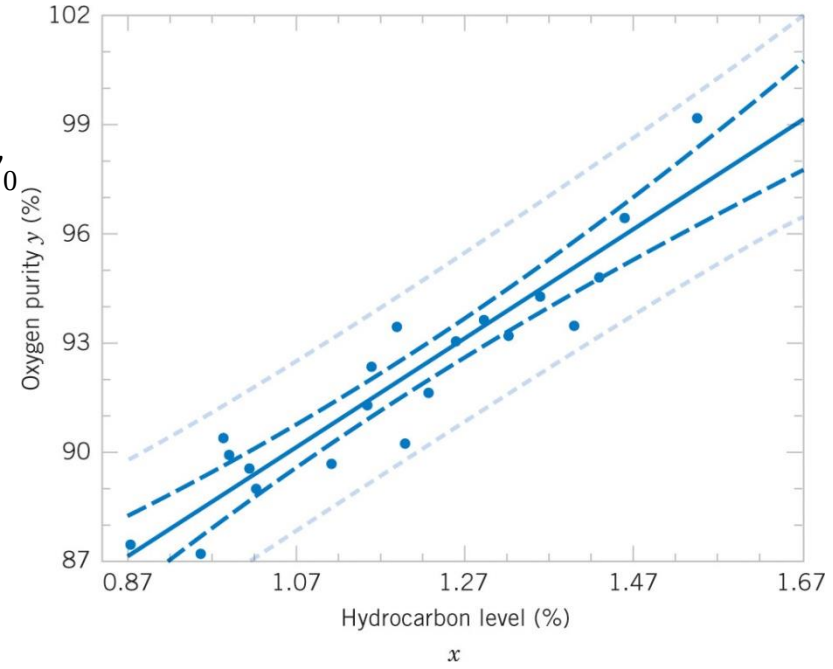
Prediction of New Observations

- Example: Chemical distillation process
  - suppose we use the data in previous example and find a 95% prediction interval on the next observation of oxygen purity a t x_0=1.00%.

$$89.23 - 2.101\sqrt{1.18[1+\frac{1}{20}+\frac{(1-1.196)^2}{0.68088}]} \le y_0$$

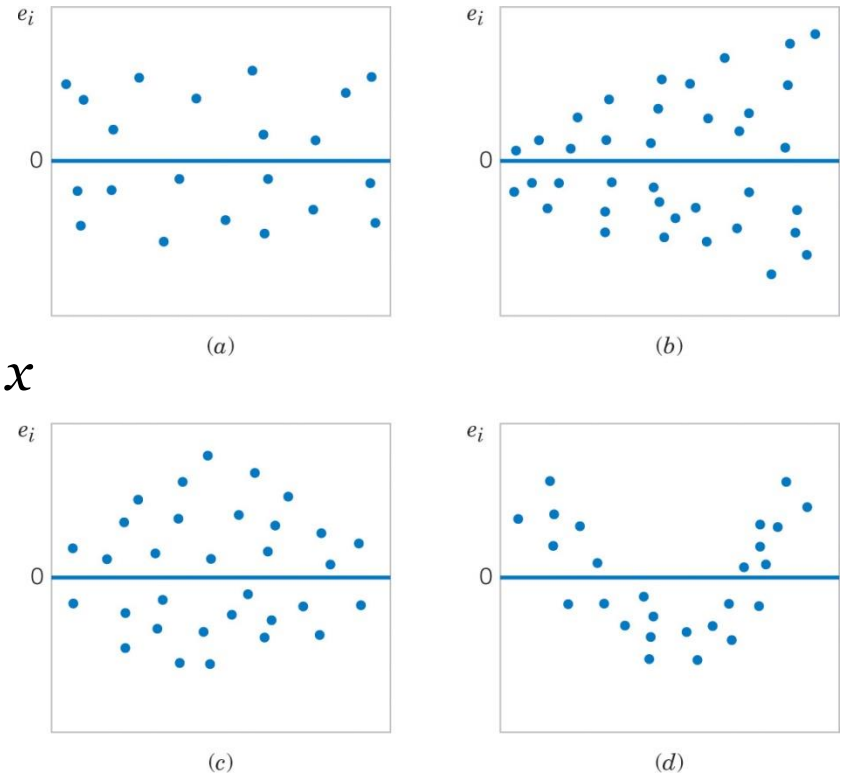
$$\le 89.23 - 2.101\sqrt{1.18[1+\frac{1}{20}+\frac{(1-1.196)^2}{0.68088}]}$$

$$86.83 \le y_0 \le 91.63$$



Adequacy of the Regression Model

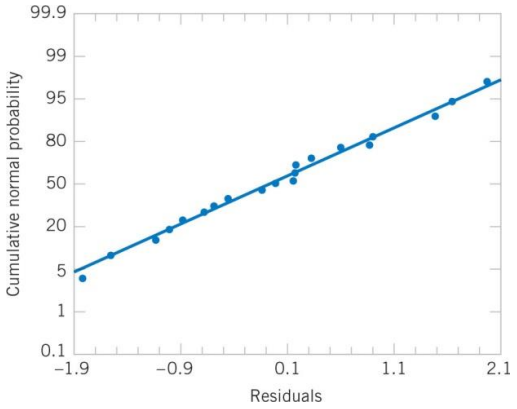
- Our assumption in LSE
  - Errors are uncorrelated with mean zero and constant varianc e
  - Normally distributed for tests of hypothesis
- Residual analysis
  - Residual $e_i = y_i - \hat{y}_i$
- Residual plot
  - In time sequence
  - Against the $\hat{y}_i$
  - Against the independent x

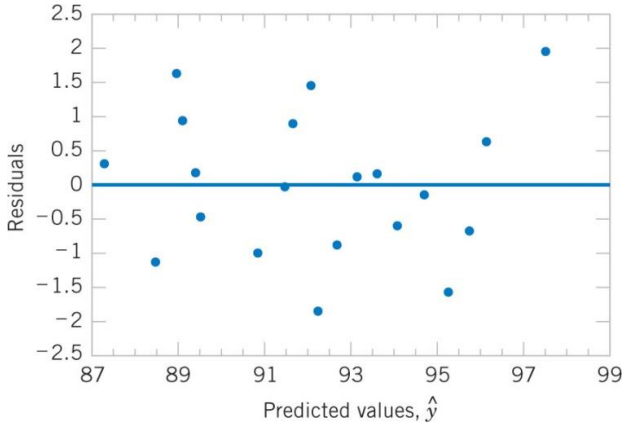


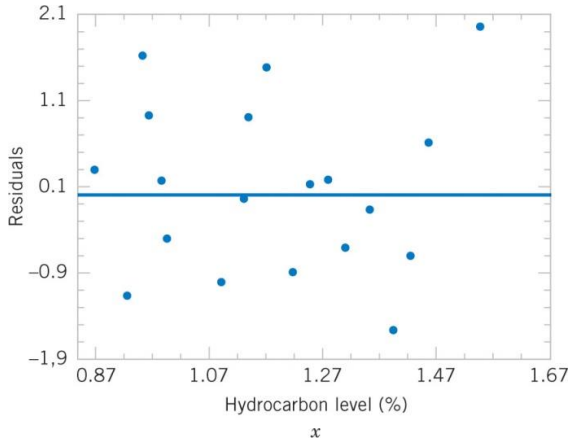
Adequacy of the Regression Model

- Example: chemical distillation process

	Hydrocarbon Level, x	Oxygen Purity, y	Predicted Value, $\hat{y}$	Residual $e = y - \hat{y}$		Hydrocarbon Level, x	Oxygen Purity, y	Predicted Value, $\hat{y}$	Residual $e = y - \hat{y}$
1	0.99	90.01	89.081	0.929	11	1.19	93.54	92.071	1.469
2	1.02	89.05	89.530	−0.480	12	1.15	92.52	91.473	1.047
3	1.15	91.43	91.473	−0.043	13	0.98	90.56	88.932	1.628
4	1.29	93.74	93.566	0.174	14	1.01	89.54	89.380	0.160
5	1.46	96.73	96.107	0.623	15	1.11	89.85	90.875	−1.025
6	1.36	94.45	94.612	−0.162	16	1.20	90.39	92.220	−1.830
7	0.87	87.59	87.288	0.302	17	1.26	93.25	93.117	0.133
8	1.23	91.77	92.669	−0.899	18	1.32	93.41	94.014	−0.604
9	1.55	99.42	97.452	1.968	19	1.43	94.98	95.658	−0.678
10	1.40	93.65	95.210	−1.560	20	0.95	87.33	88.483	−1.153







ANOVA to Test Significance of Regression

- Analysis of variance identity

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y})^2$$

  - Error sum of squares: SS_E
  - Regression sum of squares: SS_R
  - Total corrected sum of squares: SS_T
- Test for significance of regression

$$F_0=\frac{SS_R/1}{SS_E/(n-2)}=\frac{MS_R}{MS_E}$$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	$SS_R=\hat{\beta}_1 S_{xy}$	1	MS_R	MS_R/MS_E
Error	$SS_E=SS_T-\hat{\beta}_1 S_{xy}$	$n-2$	MS_E	
Total	SS_T	$n-1$		

Note that $MS_E=\hat{\sigma}^2$.

ANOVA to Test Significance of Regression

- Example: Chemical distillation process
  - Recall previous oxygen purity data model

$$SS_R = \hat{\beta}_1 S_{xy} = (14.947)10.17744 = 152.13$$

$$SS_E = SS_T - SS_R = 173.38 - 152.13 = 21.25$$

  - Our test statistic

$$f_0 = \frac{MS_R}{MS_E} = \frac{152.13}{21.25/18} = 128.86$$

Adequacy of the Regression Model

- Coefficient of determination (R^2)

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

- Correlation
  - Assume the observations (X_i, Y_i) are jointly distributed rando m variables
  - Correlation coefficient

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

  - Conditional distribution (bivariate normal)

$$f_{Y|x}(y) = \frac{1}{\sqrt{2\pi}\sigma_{Y|x}} \exp\left[-\frac{1}{2}\left(\frac{y - \beta_0 - \beta_1 x}{\sigma_{Y|x}}\right)^2 \right]$$

where

$$\beta_0 = \mu_Y - \mu_X \rho \frac{\sigma_Y}{\sigma_X}, \quad \beta_1 = \frac{\sigma_Y}{\sigma_X}\rho, \quad \sigma^2_{Y|x} = \sigma^2_Y(1-\rho^2)$$

Adequacy of the Regression Model

- Using MLE to estimate parameter β_0 and β_1

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i (X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{S_{XY}}{S_{XX}}$$

- Sample correlation coefficient

$$R = \frac{\sum_{i=1}^{n} Y_i (X_i - \bar{X})}{[\sum_{i=1}^{n} (X_i - \bar{X})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2]^{1/2}} = \frac{S_{XY}}{(S_{XX} SS_T)^{1/2}}$$

$$\hat{\beta}_1 = (\frac{SS_T}{S_{XX}})^{1/2} R$$

Adequacy of the Regression Model

- Test of coefficient of correlation
  - If Y and X are correlated

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

  - Test statistic

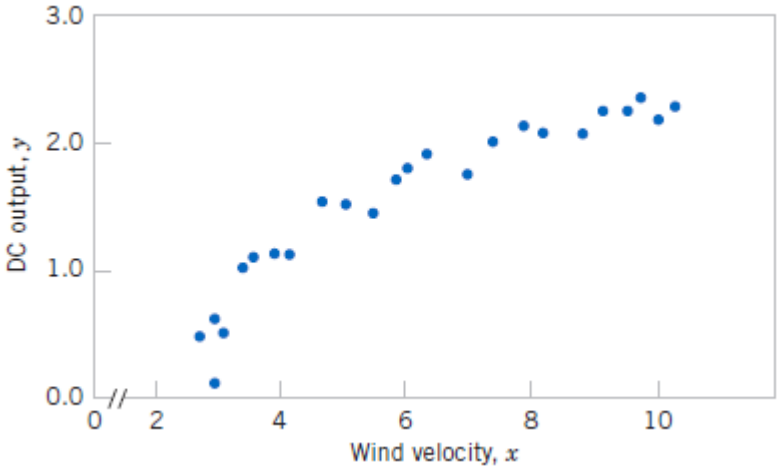
$$T_0 = \frac{R(\sqrt{n-2})}{\sqrt{1-R^2}} \sim t_{n-2}$$

  - Reject null hypothesis if

$$|t_0| > t_{\alpha/2,n-2}$$

Transformations

- If the true regression function is nonlinear
  - Transformation may work
- Example
  - A research engineer is investigating the use of a windmill to generate electricity and has collected data on the DC output from this windmill and the corresponding wind velocity.



Observation Number, i	Wind Velocity (mph), x_i	DC Output, y_i
4	2.70	0.500
5	10.00	2.236
6	9.70	2.386
7	9.55	2.294
8	3.05	0.558
9	8.15	2.166
10	6.20	1.866
11	2.90	0.653
12	6.35	1.930
13	4.60	1.562
14	5.80	1.737
15	7.40	2.088
16	3.60	1.137
17	7.85	2.179
18	8.80	2.112
19	7.00	1.800
20	5.45	1.501
21	9.10	2.303
22	10.20	2.310
23	4.10	1.194
24	3.95	1.144
25	2.45	0.123

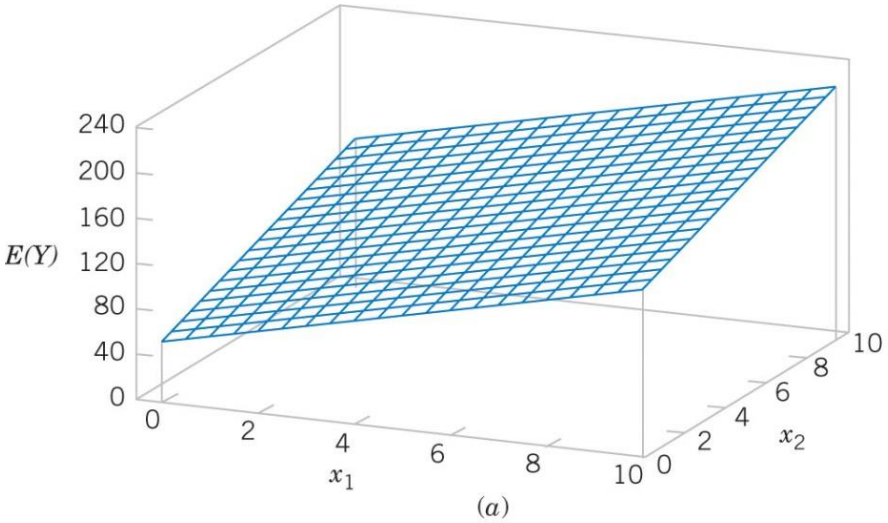
Multiple Linear Regression

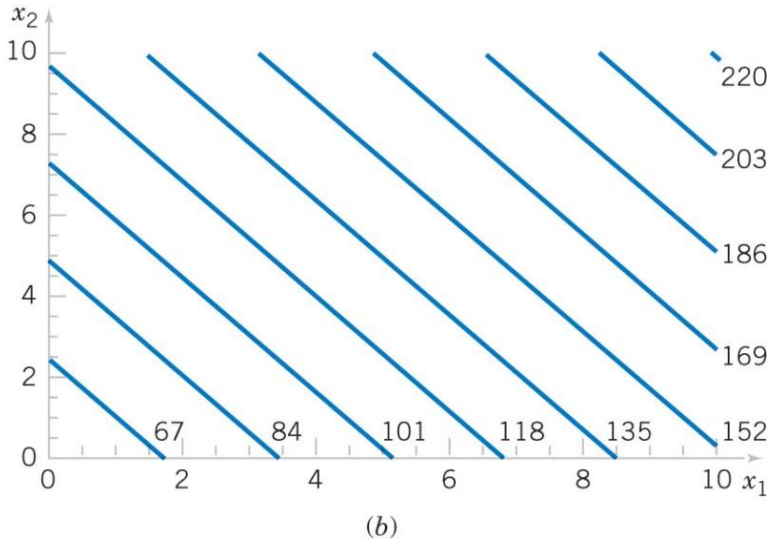
- A regression model that contains more than one regre ssor variable

$$Y = \beta_0 + \beta_1 x + \beta_2 x_2 + \varepsilon$$

  - Example:

$$E(Y) = 50 + 10x_1 + 7x_2$$





Multiple Linear Regression

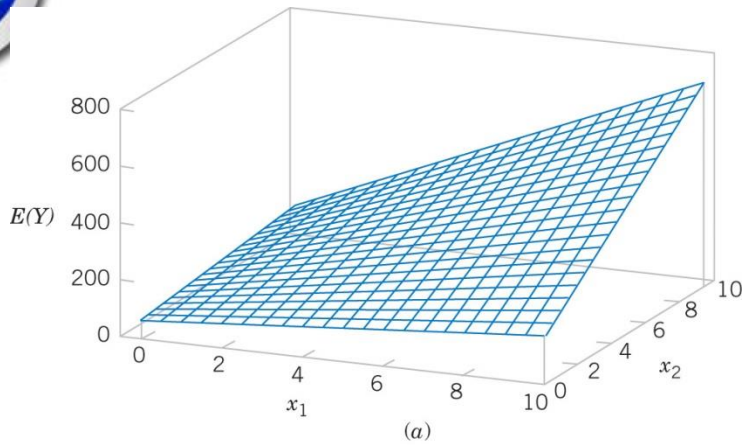
- The model that Y may be related to k independent vari ables

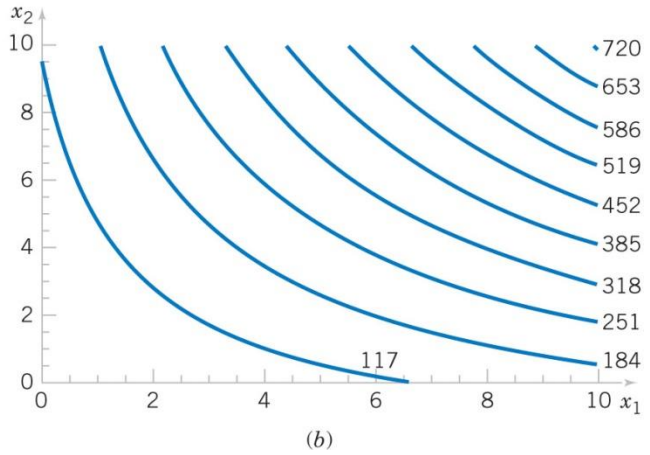
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

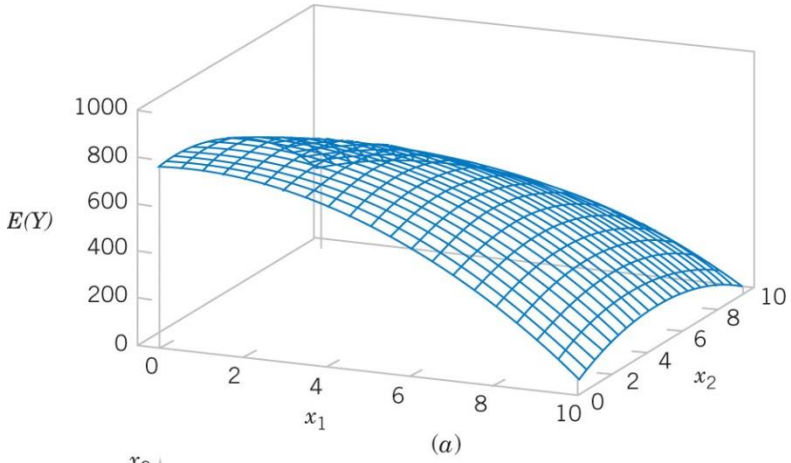
- If system structure is complex
  - e.g. : polynomial $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_k x^3 + \varepsilon$
- Models with interactions

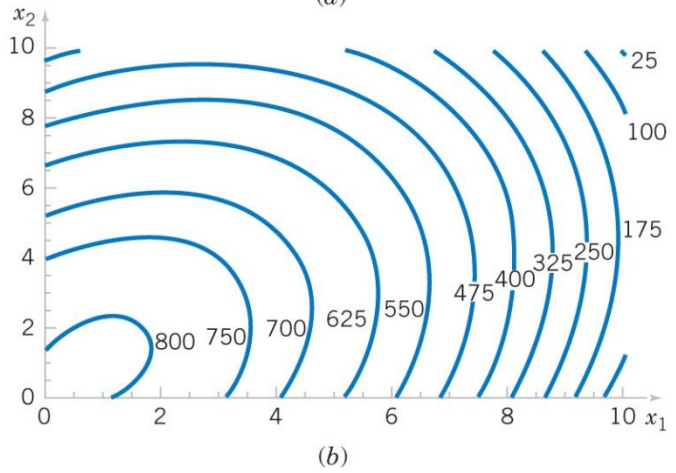
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon$$

- Any regression model that is linear in parameter is a li near regression model, regardless of the shape of the s urface it generates









$$E(Y) = 50 + 10x_1 + 7x_3 + 5x_1x_3$$

$$E(Y) = 800 + 10x_1 + 7x_2 - 8.5x_1^2 - 5x_2^2 + 4x_1x_2$$

LSE of the Parameters

- Suppose that the observations are $(x_{i1}, x_{i2}, \ldots x_{ik}, y_i)$, $n > k$

$$y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

$$= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i$$

- The least squares function

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij})^2$$

- Obtain the estimate of the parameters by minimizing L

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k,} = -2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij}) = 0$$

$$\left.\frac{\partial L}{\partial \beta_j}\right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k,} = -2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij}) x_{ij} = 0 \qquad j = 1,2,\ldots, k$$

- Least squares normal equations

$$
\begin{aligned}
n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} \quad &+\hat{\beta}_2 \sum_{i=1}^{n} x_{i2} \quad +\ldots+ \hat{\beta}_k \sum_{i=1}^{n} x_{ik} \quad = \sum_{i=1}^{n} y_i \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 \quad &+\hat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} +\ldots+ \hat{\beta}_k \sum_{i=1}^{n} x_{i1}x_{ik} = \sum_{i=1}^{n} x_{i1}y_i \\
\vdots \qquad\qquad \vdots \qquad\qquad &\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik}x_{i1} &+\hat{\beta}_2 \sum_{i=1}^{n} x_{ik}x_{i2} +\ldots+ \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 \quad = \sum_{i=1}^{n} x_{ik}y_i
\end{aligned}
$$

- Matrix approach to multiple linear regression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- The least square function

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

- Normal equations

$$\mathbf{X}^T \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T \mathbf{y} \Rightarrow \hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

- Estimating σ^2

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p} = \frac{SS_E}{n-p}$$

Properties of the Least Squares Estimators

- The least squares estimators are unbiased estimators

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

- Variance of estimators

$$\text{cov}(\hat{\boldsymbol{\beta}}) = E\{[\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})][\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})]^T\} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

- Define $C = (X^T X)^{-1}$

$$V(\hat{\beta}_j) = \sigma^2 C_{jj}$$

$$\text{cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 C_{ij}$$

Test for Significance of Regression

- Hypothesis for ANOVA test

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$
$$H_1 : \beta_j \neq 0$$

- Test statistic

$$F_0 = \frac{SS_R / k}{SS_E /(n-p)} = \frac{MS_R}{MS_E}$$

- Reject null if $f_0 > f_{\alpha,k,n-p}$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	SS_R	k	MS_R	MS_R/MS_E
Error or residual	SS_E	$n - p$	MS_E	
Total	SS_T	$n - 1$		

Test for Significance of Regression

- R^2 adjustment

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

  - Always increasing as the number of the variable increas es

- Adjusted R^2

$$R^2_{adj} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)}$$

Test for Significance of Regression

- Tests on individual regression coefficients ($\varepsilon \sim$ NIID)
- Hypothesis

$$H_0 : \beta_j = \beta_{j0}$$
$$H_1 : \beta_j \neq \beta_{j0}$$

- Test statistic

$$T_0 = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\sigma^2 C_{jj}}}$$

- Reject null if $|t_0| > t_{\alpha/2,n-p}$

- Example : Wire bond strength

Example

- Three variables that were collected in an observational study in a semiconductor manufacturing plant. In this plant, the finished semiconductor is wire bonded to a frame. The variables reported are pull strength (a measure of the amount of force required to break the bond), the wire length, and the height of the die. We would like to find a model relating pull strength to wire length and die height. Unfortunately, there is no physical mechanism that we can easily apply here, so it doesn't seem likely that a mechanistic modeling approach will be successful.

Observation Number	Pull Strength y	Wire Length x_1	Die Height x_2
1	9.95	2	50
2	24.45	8	110
3	31.75	11	120
4	35.00	10	550
5	25.02	8	295
6	16.86	4	200
7	14.38	2	375
8	9.60	2	52
9	24.35	9	100
10	27.50	8	300
11	17.08	4	412
12	37.00	11	400
13	41.95	12	500
14	11.66	2	360
15	21.65	4	205
16	17.89	4	400
17	69.00	20	600
18	10.30	1	585
19	34.93	10	540
20	46.59	15	250
21	44.88	15	290
22	54.12	16	510
23	56.63	17	590
24	22.13	6	100
25	21.15	5	400

CIs in Multiple Linear Regression

- CI on a regression coefficient

$$\hat{\beta}_j - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \le \beta_j \le \hat{\beta}_j + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}}$$

- CI on the mean response given a particular point

$$\hat{\mu}_{Y|x_0} - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 \mathbf{x_0}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x_0}} \le \mu_{Y|x_0} \le \hat{\mu}_{Y|x_0} + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 \mathbf{x_0}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x_0}}$$

- Prediction interval

$$\hat{y}_0 - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2[1+\mathbf{x_0}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x_0}]} \le Y_0 \le \hat{y}_0 + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2[1+\mathbf{x_0}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x_0}]}$$

Polynomial Regression Model

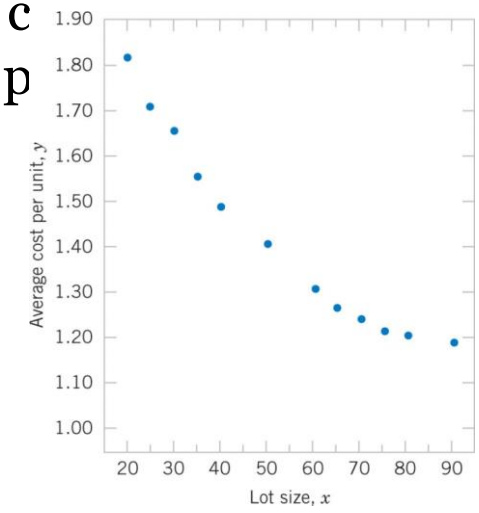
- Polynomial will be considered if the response is curvilinear
  - E.g. second degree polynomial

$$Y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \varepsilon$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon$$

- Example:
  - Sidewall panels for the interior of an airplane are formed in a 1500-ton press. The unit man ufacturing cost varies with the production lot size. The data shown below give the average c ost per unit (in hundreds of dollars) for this p roduct (y) and the production lot size (x)

y	1.81	1.70	1.65	1.55	1.48	1.40
x	20	25	30	35	40	50
y	1.30	1.26	1.24	1.21	1.20	1.18
x	60	65	70	75	80	90



Categorical Regression

- If predictors are not measured on a numerical scale
- Example:
  - A mechanical engineer is investigating the surface finish of metal parts produced on a lathe and its relationship t o the speed (in revolutions per minute) of the lathe.
  - Use dummy variables to indicate categorical variable

Observation Number, i	Surface Finish y_i	RPM	Type of Cutting Tool	Observation Number, i	Surface Finish y_i	RPM	Type of Cutting Tool
1	45.44	225	302	11	33.50	224	416
2	42.03	200	302	12	31.23	212	416
3	50.10	250	302	13	37.52	248	416
4	48.75	245	302	14	37.13	260	416
5	47.92	235	302	15	34.70	243	416
6	47.79	237	302	16	33.92	238	416
7	52.26	265	302	17	32.13	224	416
8	50.52	259	302	18	35.47	251	416
9	45.58	221	302	19	33.49	232	416
10	44.78	218	302	20	32.29	216	416

Selection of Variables

- Screening the candidate variables to obtain a regression model that contains the best subset of regressor variables
  - Contain enough regressor variables (precise)
  - To keep model maintenance costs to a minimum (robust)
- Criterion
  - R^2 and R^2_{adj}
  - C_p

$$C_P = \frac{SS_E}{\hat{\sigma}^2} - n + 2p$$

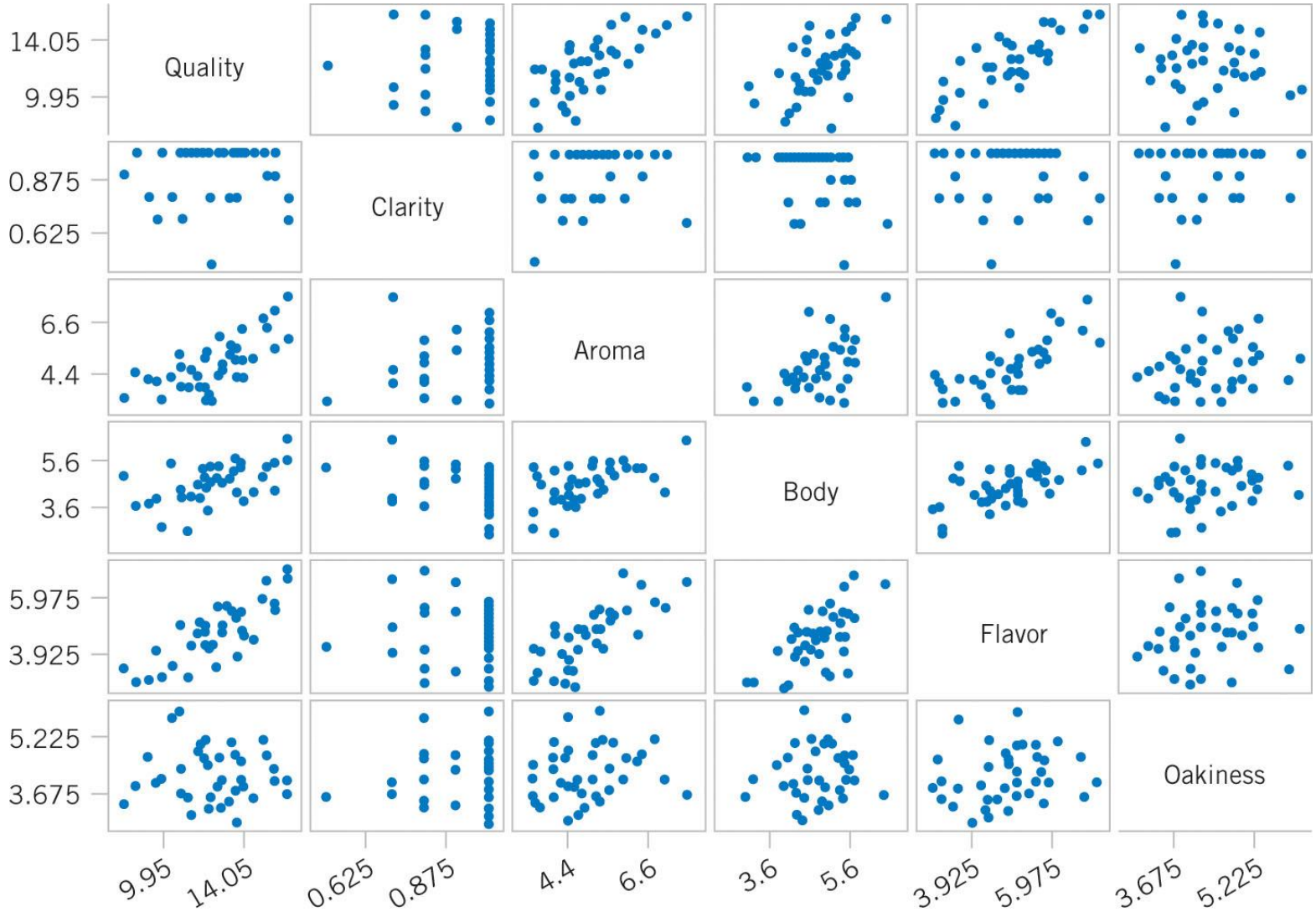
  - Prediction error sum of squares (PRESS)

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2 = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}}\right)^2$$

Example

- 38 brands of pinot noir wine are tasted and evaluated. The response variable is y = quality, and we wish to find the "best" regression equation that relates quality to the other five parameters

	x_1 Clarity	x_2 Aroma	x_3 Body	x_4 Flavor	x_5 Oakiness	y Quality
1	1.0	3.3	2.8	3.1	4.1	9.8
2	1.0	4.4	4.9	3.5	3.9	12.6
3	1.0	3.9	5.3	4.8	4.7	11.9
4	1.0	3.9	2.6	3.1	3.6	11.1
5	1.0	5.6	5.1	5.5	5.1	13.3
6	1.0	4.6	4.7	5.0	4.1	12.8
7	1.0	4.8	4.8	4.8	3.3	12.8
8	1.0	5.3	4.5	4.3	5.2	12.0
9	1.0	4.3	4.3	3.9	2.9	13.6
10	1.0	4.3	3.9	4.7	3.9	13.9
11	1.0	5.1	4.3	4.5	3.6	14.4
12	0.5	3.3	5.4	4.3	3.6	12.3
13	0.8	5.9	5.7	7.0	4.1	16.1
14	0.7	7.7	6.6	6.7	3.7	16.1
15	1.0	7.1	4.4	5.8	4.1	15.5
16	0.9	5.5	5.6	5.6	4.4	15.5
17	1.0	6.3	5.4	4.8	4.6	13.8
18	1.0	5.0	5.5	5.5	4.1	13.8
19	1.0	4.6	4.1	4.3	3.1	11.3
20	0.9	3.4	5.0	3.4	3.4	7.9
21	0.9	6.4	5.4	6.6	4.8	15.1
22	1.0	5.5	5.3	5.3	3.8	13.5
23	0.7	4.7	4.1	5.0	3.7	10.8
24	0.7	4.1	4.0	4.1	4.0	9.5
25	1.0	6.0	5.4	5.7	4.7	12.7
26	1.0	4.3	4.6	4.7	4.9	11.6
27	1.0	3.9	4.0	5.1	5.1	11.7
28	1.0	5.1	4.9	5.0	5.1	11.9
29	1.0	3.9	4.4	5.0	4.4	10.8
30	1.0	4.5	3.7	2.9	3.9	8.5
31	1.0	5.2	4.3	5.0	6.0	10.7
32	0.8	4.2	3.8	3.0	4.7	9.1
33	1.0	3.3	3.5	4.3	4.5	12.1
34	1.0	6.8	5.0	6.0	5.2	14.9
35	0.8	5.0	5.7	5.5	4.8	13.5
36	0.8	3.5	4.7	4.2	3.3	12.2
37	0.8	4.3	5.5	3.5	5.8	10.3
38	0.8	5.2	4.8	5.7	3.5	13.2


Quality
Clarity
Aroma
Body
Flavor
Oakiness
14.05
9.95
0.875
0.625
6.6
4.4
5.6
3.6
5.975
3.925
5.225
3.675
9.95
14.05
0.625
0.875
4.4
6.6
3.6
5.6
3.925
5.975
3.675
5.225

Stepwise Regression

- Iteratively constructs a sequence of regression models by adding or removing variables at each step by F test
  - f_{in}: the F value for adding a variable to the model
  - f_{out}: the F value for removing a variable from the model
  - Procedures
    - Start from one-variable model in which the predictor has the h ighest correlation with Y
    - The remaining K-1 candidate variables are examined by F-stati stic

$$F_j = \frac{SS_R(\beta_j \mid \beta_1, \beta_0)}{MS_E(x_j, x_1)}$$

    - The variable with maximum F is added to the equation if $f_j > f_{in}$
    - Examine x_1 by F-statistic

$$F_j = \frac{SS_R(\beta_1 \mid \beta_2, \beta_0)}{MS_E(x_1, x_2)}$$

    - The variable x_1 is removed if the calculated value $f_1 < f_{out}$

Summary

- Introduction
- Simple linear regression
  - Parameter estimation $Y = \beta_0 + \beta_1 x + \varepsilon$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}$$

  - Hypothesis tests on parameters

$$T_0 = \frac{\hat{\beta}_1 - \hat{\beta}_{1,0}}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \qquad T_0 = \frac{\hat{\beta}_0 - \hat{\beta}_{0,0}}{\sqrt{\hat{\sigma}^2 [\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}]}}$$

Summary

- Simple linear regression $Y = \beta_0 + \beta_1 x + \varepsilon$
  - Confidence intervals on parameters

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]} \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}$$

$$\hat{\mu}_{Y|x_0} - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}]} \le \mu_{Y|x_0} \le \hat{\mu}_{Y|x_0} + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}]}$$

  - Prediction of new observations

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

$$\hat{y}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}]} \le Y_0 \le \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}]}$$

Summary

- Simple linear regression $Y = \beta_0 + \beta_1 x + \varepsilon$
  - Model adequacy

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

- Multiple linear regression

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

  - Parameter estimation

$$\mathbf{X}^T \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T \mathbf{y} \Rightarrow \hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

  - Hypothesis tests on parameters

$$F_0 = \frac{SS_R / k}{SS_E / (n-p)} = \frac{MS_R}{MS_E} \qquad T_0 = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\sigma^2 C_{jj}}}$$

Summary

- Multiple linear regression
  - Confidence intervals

$$\hat{\beta}_j - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \le \beta_j \le \hat{\beta}_j + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}}$$

$$\hat{\mu}_{Y|x_0} - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 \mathbf{x_0}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x_0}} \le \mu_{Y|x_0} \le \hat{\mu}_{Y|x_0} + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 \mathbf{x_0}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x_0}}$$

$$\hat{y}_0 - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2[1+\mathbf{x_0}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x_0}]} \le Y_0 \le \hat{y}_0 + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2[1+\mathbf{x_0}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x_0}]}$$

  - Polynomial regression model
  - Categorical regression
  - Variable selection
  - Stepwise regression