Medical statistics

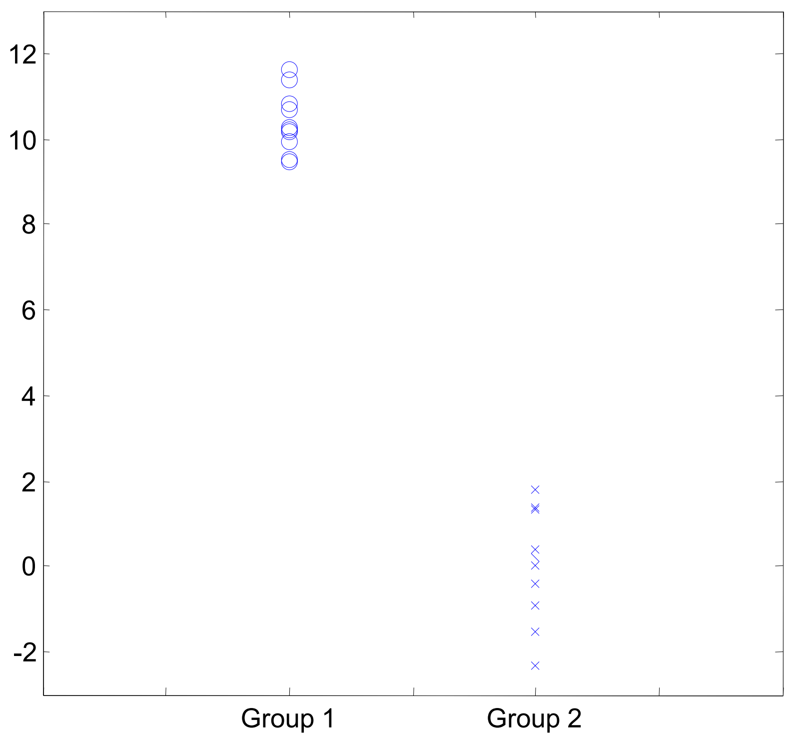
Ref: Montgomery DC, Runger GC, and Hubele NF, "Engineering statistics", 5th ed., 2010

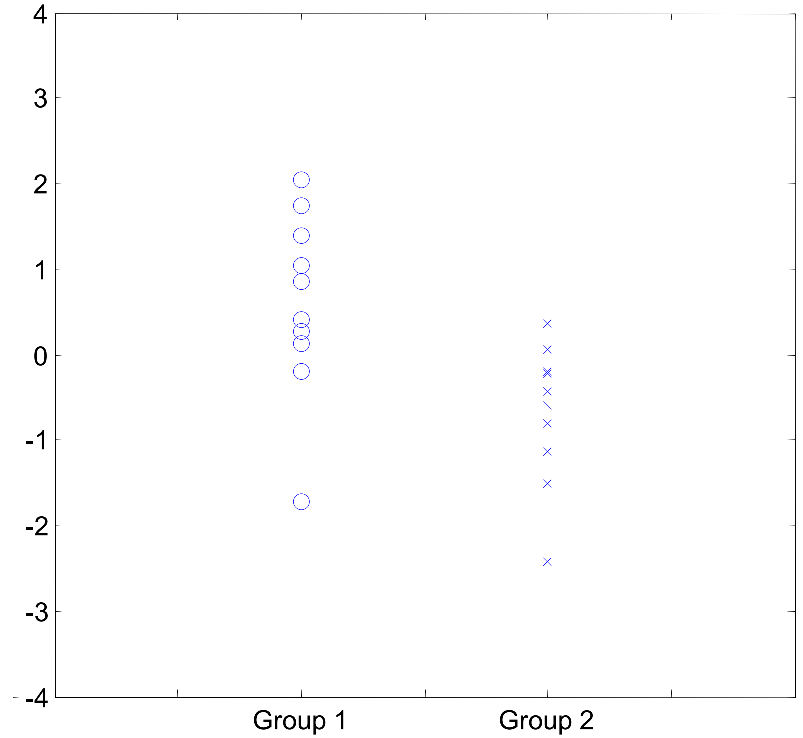
莊子肇 副教授
中山電機系

Why statistics?

- Example: Any regional volume change of gray matter or white matter on patients with attention-deficit hyperactivity disorder (ADHD)?
  - WM/GM separation
  - Need inter-subject comparison
  - Recall the methods of image registration
  - Then?

Why Statistics?





Why statistics?

- There is always **uncertainty** in real world.

- Another example: Suppose that one of your friends, seeing that you are struggling with your course works, suggested that you quit the school, citing that “Mr. XXX made billions of dollars without even finishing elementary school!”

  Would you follow the advice? Variation and risk?

Outline

- Data summary and presentation
- Random variables and probability
- Decision making
  - Parameter estimation & hypothesis testing
  - Z-test and T-test
  - One sample and two samples
  - Analysis of variance (ANOVA)

Population and sample

- Question: What is the average body temperature of monkeys in Taiwan?

- What we (engineers) do in most engineering applications:
  - Collect (all?) data
  - Predict what other untested samples will perform
  - **Statistical inference**

Population and sample

- Question: What is the average body temperature of monkeys in Taiwan?
  - Population: all monkeys in Taiwan
  - Sample: 10 age- and gender-matched monkeys from each county
  - How close is the sample mean to the population mean?
  - How close is the sample variation to the population variation?

Statistics

- Statistics is the science of collecting, summarizing, presenting, and interpreting data.
  - Reliable data
  - Consistent analysis
  - Be as simple as possible and make sense!

Data summary

- Sample mean ($\bar{x}$)
- Population mean (μ)
  - The sample mean is a reasonable estimate of the population mean.

Data summary

- Sample variance
- Sample standard deviation (s)

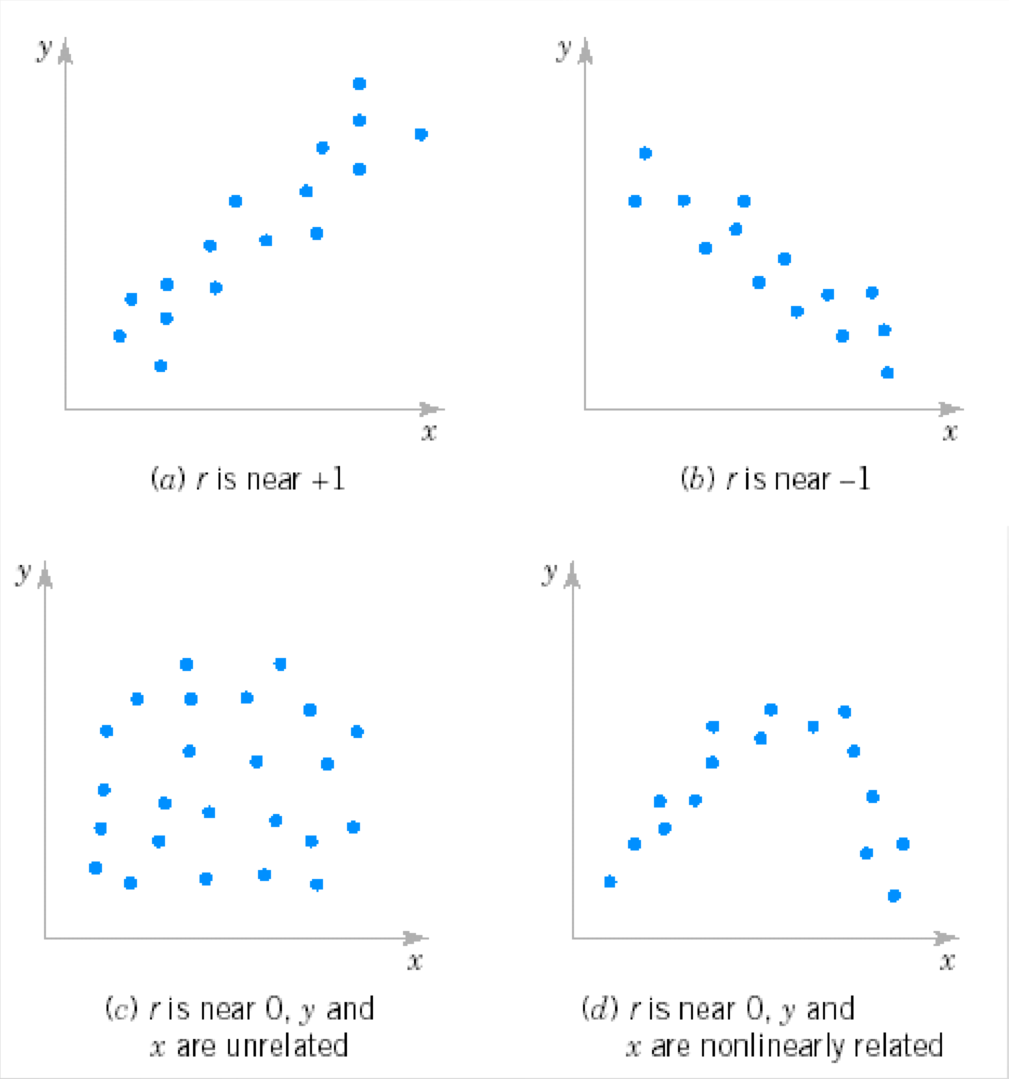
- Population variance
- Population standard deviation (σ)

  - The sample variance is a reasonable estimate of the population variance.

Data summary

- Coefficient of variation
- Sample correlation coefficient (*r*)

Scatter diagram of multivariate



Data display: Stem-and-Leaf diagram

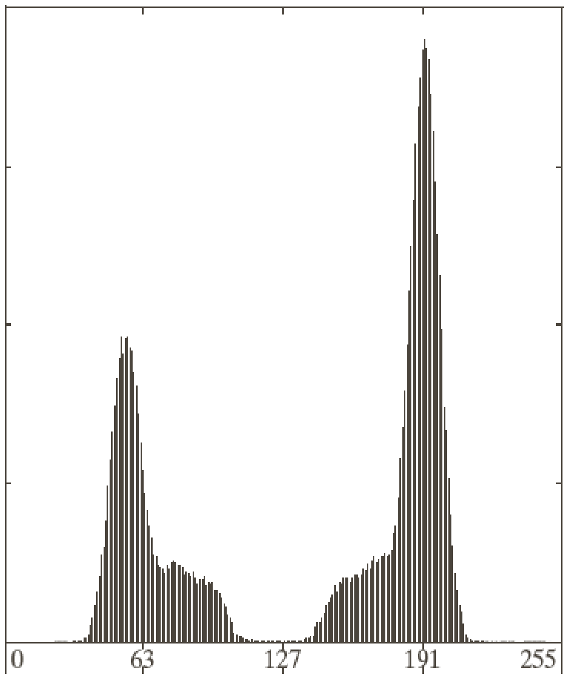
Barry Bonds

Year	Age	OBP	SLG	OPS
1986	22	.330	.416	.746
1987	23	.329	.492	.821
1988	24	.368	.491	.859
1989	25	.351	.426	.777
1990	26	.406	.565	.971
1991	27	.410	.514	.924
1992	28	.456	.624	1.080
1993	29	.458	.677	1.135
1994	30	.426	.647	1.073
1995	31	.431	.577	1.008
1996	32	.461	.615	1.076
1997	33	.446	.585	1.031
1998	34	.438	.609	1.047
1999	35	.389	.617	1.006
2000	36	.450	.707	1.157
2001	37	.515	.863	1.378
2002	38	.582	.799	1.381
2003	39	.529	.749	1.278
2004	40	.609	.812	1.422
2005	41	.404	.667	1.071
2006	42	.454	.545	.999
2007	43	.480	.565	1.045
	22 (avg)	.444	.607	1.051

Stemplot of Bonds' OPS (*N* = 22)

Stem	*Leaf*
.7	5 8
.8	2 6
.9	2 7
1.0	0 1 1 3 5 5 7 7 8 8
1.1	4 6
1.2	8
1.3	8 8
1.4	2

Data display: Histogram



- More compact

Data display: Pareto chart

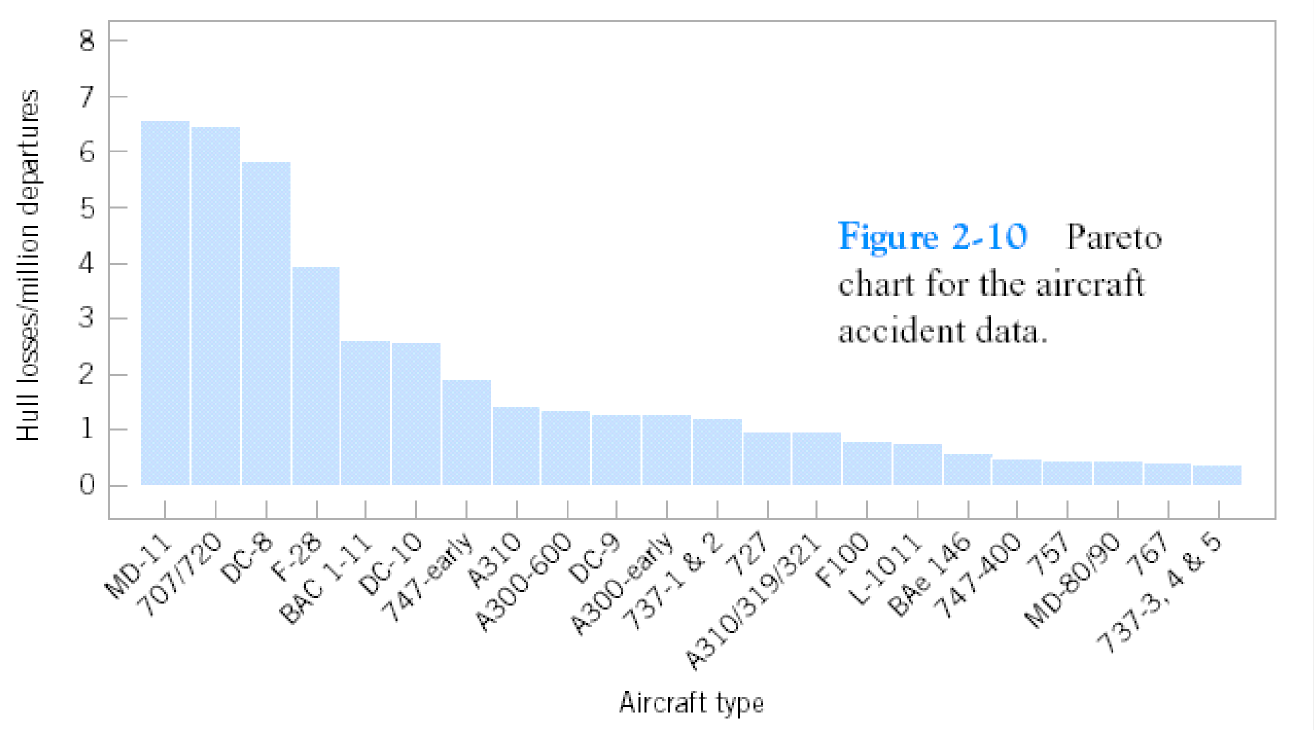
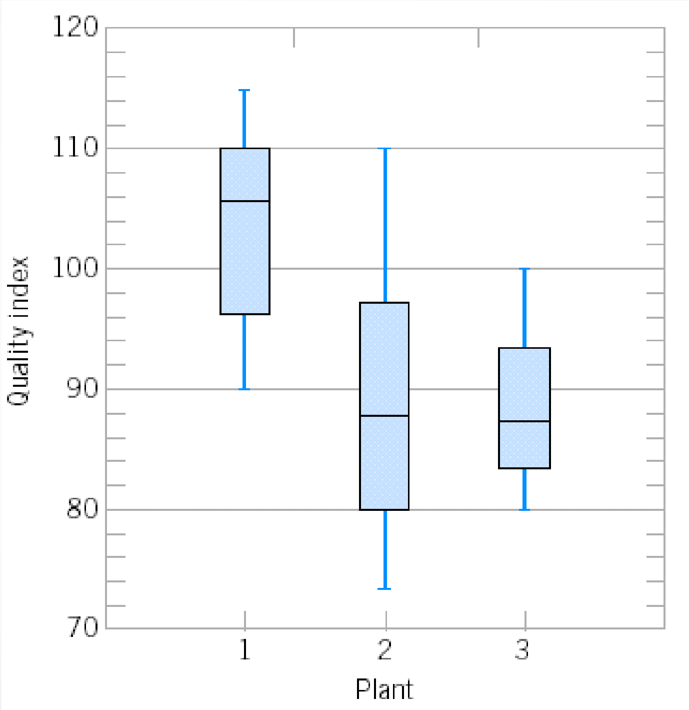


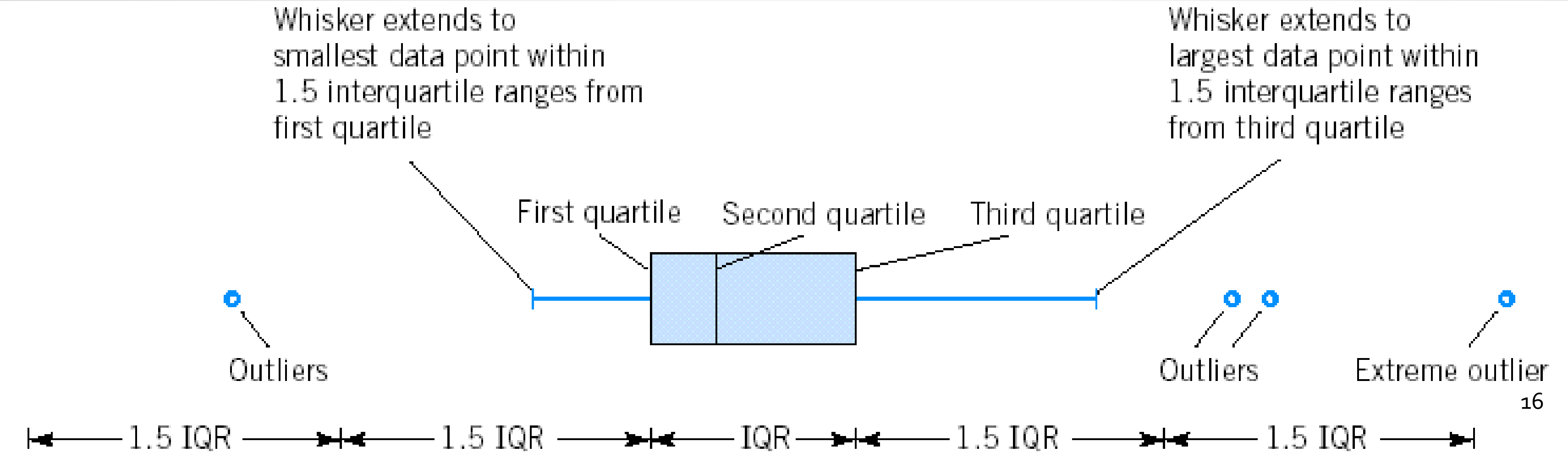
Figure 2-10 Pareto chart for the aircraft accident data.

- Widely used in quality and process improvement studies

Data display: Box plot

- Main features
  - Q1, Median (Q2), Q3
  - IQR: interquartile range
  - Whisker
  - Outlier





Random variables and probability distributions

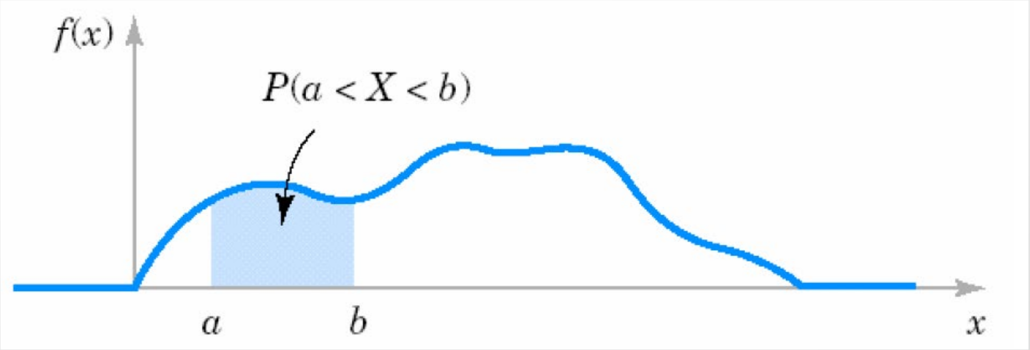
Random variable

- A **random variable** is a numerical variable whose measured value can change from one replicate of the experiment to another.
  - Example: body temperature, electric current, number of transmitted bits received in error
- **Probability**: the likelihood that particular values occur

Probability

- Probability density function (PDF, $f(x)$)
  - Properties of the PDF*

$P(\underbrace{a < X < b}_{\text{event}}) =$



  - Approximated by histogram.

Probability

- Cumulative distribution function (CDF, $F(x)$)
  - $F(x) = P(X \leqq x)$
  - $P(a < X < b) = ?$

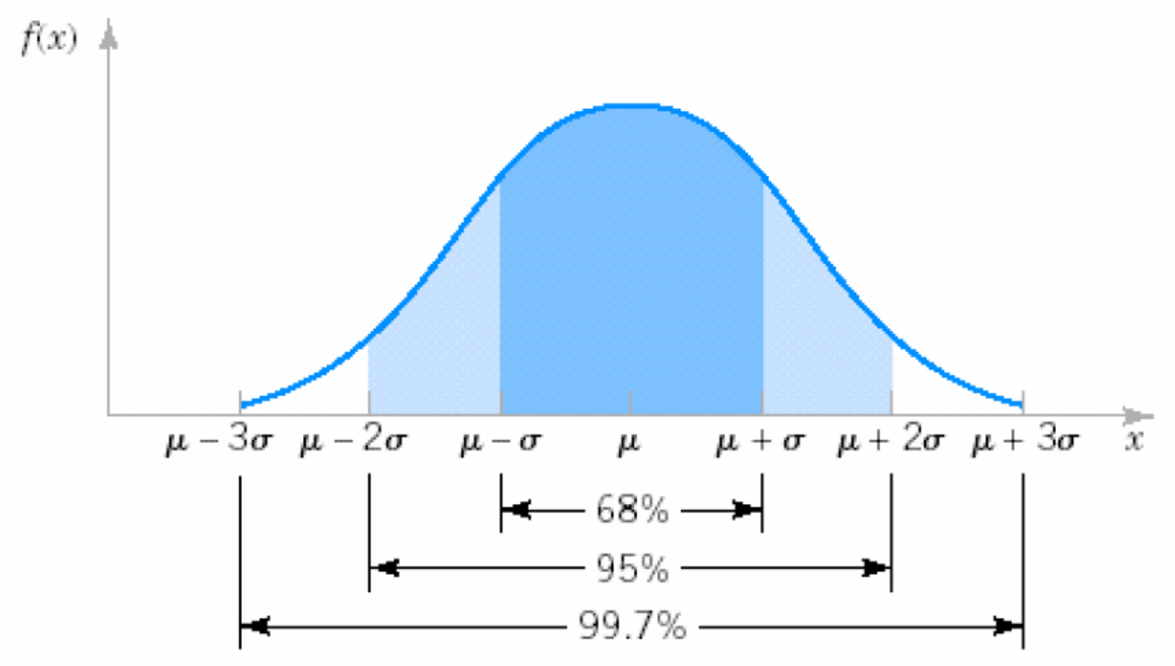
Continuous random variables

- Suppose X is a continuous random variable…
  - Mean (expected value, E(x))
  - Variance (V(x))

Normal distribution

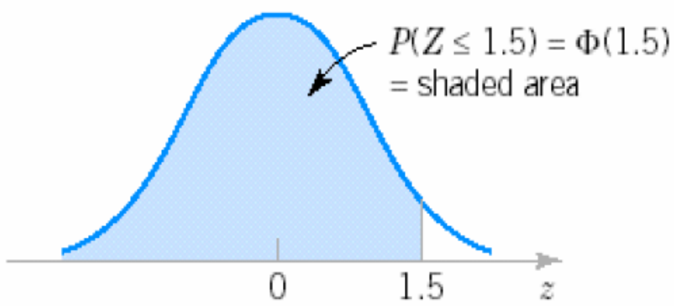
- Although distributions can have different shapes, the most widely used model for a random variable (e.g. blood pressure, height) is **normal distribution**.
  - Also referred to as a Gaussian distribution.
  - $f(x) = ?$

Normal distribution



- Standard normal distribution (Z)*
  - A normal random distribution with μ = 0 and σ^2=1

Standard normal distribution



z	0.00	0.01	0.02	0.03
0	0.50000	0.50399	0.50398	0.51197
⋮		⋮		
1.5	0.93319	0.93448	0.93574	0.93699

Z	Área	Z	Área	Z	Área	Z	Área	Z	Área	Z	Área	Z	Área	Z	Área
-4,00	0,0000	-3,00	0,0013	-2,00	0,0228	-1,00	0,1587	0,00	0,5000	1,00	0,8413	2,00	0,9772	3,00	0,9987
-3,99	0,0000	-2,99	0,0014	-1,99	0,0233	-0,99	0,1611	0,01	0,5040	1,01	0,8438	2,01	0,9778	3,01	0,9987
-3,98	0,0000	-2,98	0,0014	-1,98	0,0239	-0,98	0,1635	0,02	0,5080	1,02	0,8461	2,02	0,9783	3,02	0,9987
-3,97	0,0000	-2,97	0,0015	-1,97	0,0244	-0,97	0,1660	0,03	0,5120	1,03	0,8485	2,03	0,9788	3,03	0,9988
-3,96	0,0000	-2,96	0,0015	-1,96	0,0250	-0,96	0,1685	0,04	0,5160	1,04	0,8508	2,04	0,9793	3,04	0,9988
-3,95	0,0000	-2,95	0,0016	-1,95	0,0256	-0,95	0,1711	0,05	0,5199	1,05	0,8531	2,05	0,9798	3,05	0,9989
-3,94	0,0000	-2,94	0,0016	-1,94	0,0262	-0,94	0,1736	0,06	0,5239	1,06	0,8554	2,06	0,9803	3,06	0,9989
-3,93	0,0000	-2,93	0,0017	-1,93	0,0268	-0,93	0,1762	0,07	0,5279	1,07	0,8577	2,07	0,9808	3,07	0,9989
-3,92	0,0000	-2,92	0,0018	-1,92	0,0274	-0,92	0,1788	0,08	0,5319	1,08	0,8599	2,08	0,9812	3,08	0,9990
-3,91	0,0000	-2,91	0,0018	-1,91	0,0281	-0,91	0,1814	0,09	0,5359	1,09	0,8621	2,09	0,9817	3,09	0,9990

CDF of a standard normal distribution

Find $P(-1.92 < Z < 1.05)$ =?

Example

- The diameter of a shaft in an optical storage drive is normally distributed with mean 0.2508 inch and standard deviation 0.0005 inch. The specifications on the shaft are 0.2500 ± 0.0013 inch. What proportion of the shaft conforms to the specification?

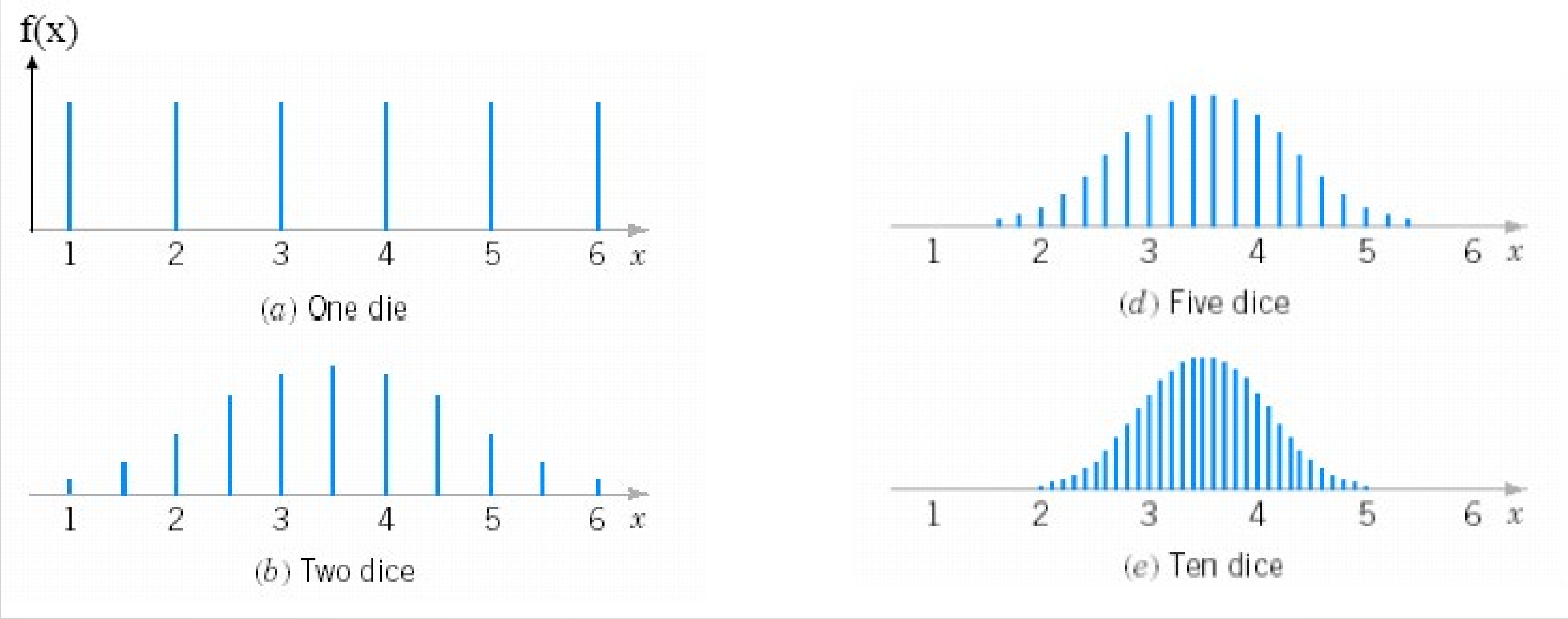
Normal distribution sampling theorem

- If a variable X is normally distributed with a mean μ and a standard deviation σ, the sampling distribution of the mean X, based on random samples of size n, will also be normally distributed and have a mean μ, and a standard deviation σ_x.

Central limit theorem

- If a variable X has a distribution ~(μ, σ^2), the sampling distribution of the mean X, based on random samples of size n, will have a mean equal to μ, and a standard distribution σ_x, and the shape will tend to be a normal distribution when $n \rightarrow \infty$.

Central limit theorem



Probability plotting

- How do we know if a normal distribution is a reasonable model for data?
- Probability plotting is a graphical method for determining whether sample data conform to a hypothesized distribution based on a subjective visual examination of the data.

Probability plotting

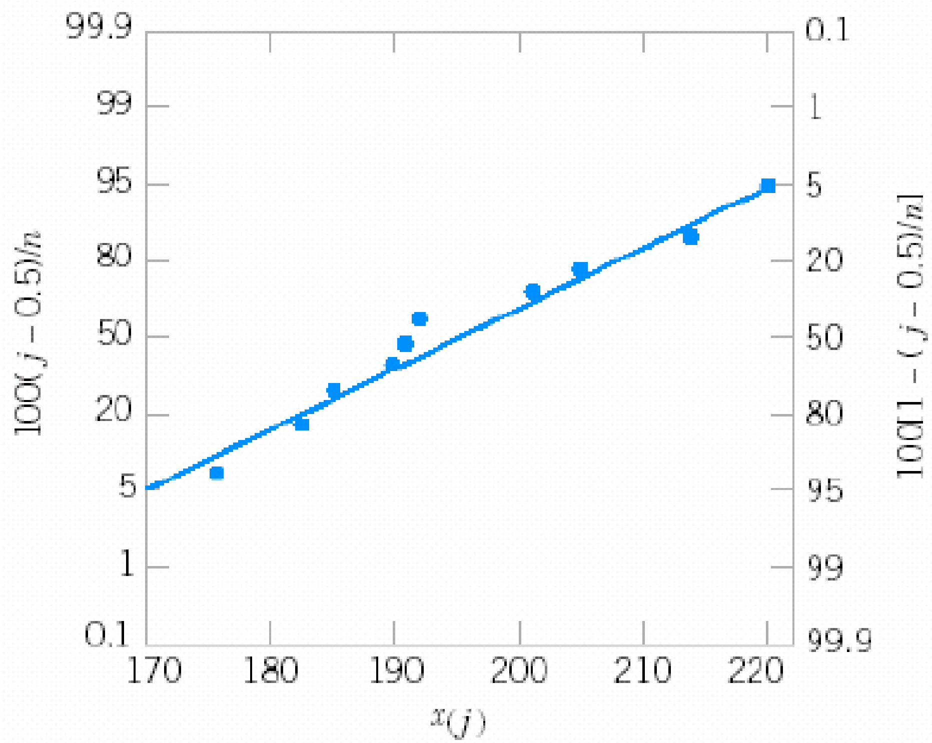


Figure 3-23 Normal probability plot for the battery life.

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

- **Rule of thumb: draw the straight line between the 25th and 75th percentile points**