Sampling Distribution

M.M. Pedram
pedram@tmu.ac.ir
Tarbiat Moallem University of Tehran
(Fall 2010)

1

Statistics/(ML, PR, Data Mining) Dictionary

Statistics	Computer Science	Meaning
estimation	learning	using data to estimate an unknown quantity
classification	supervised learning	predicting a discrete *Y* from *X*
clustering	unsupervised learning	putting data into groups
data	training sample	$(X_1,Y_1), \dots, (X_n,Y_n)$
covariates	features	the X_i's
classifier	hypothesis	a map from covariates to outcomes
Hypothesis	-	subset of a parameter space Θ
confidence interval	-	interval that contains an unknown quantity with given frequency
directed acyclic graph	Bayes net	multivariate distribution with given conditional independence relations
Bayesian inference	Bayesian inference	statistical methods for using data to update beliefs
frequentist inference	-	statistical methods with guaranteed frequency behavior
large deviation bounds	PAC learning	uniform bounds on probability of errors

2

What is Statistics?

"Statistics is a way to get information from data"

Data: Facts, especially numerical facts, collected together for reference or information.

Knowledge: High level information communicated concerning some particular fact.

Statistics is a ***tool*** for creating ***new understanding*** from a set of numbers.

3

Key Statistical Concepts

- **Population**

  a ***population*** is the group of **all** items of interest to a statistics practitioner.

  - Frequently very large; sometimes infinite.
    - E.g. all 42 million voters,

- **Sample**

  A ***sample*** is a set of data drawn from the population.

  - Potentially very large, but less than the population.
    - E.g. a sample of 2000 voters exit polled on election day.

4

Key Statistical Concepts

- **Parameter**

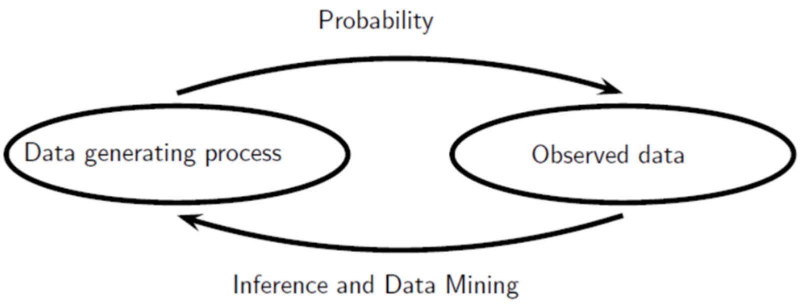
  A descriptive measure of a ***population.***

- **Statistic**

  A descriptive measure of a ***sample.***

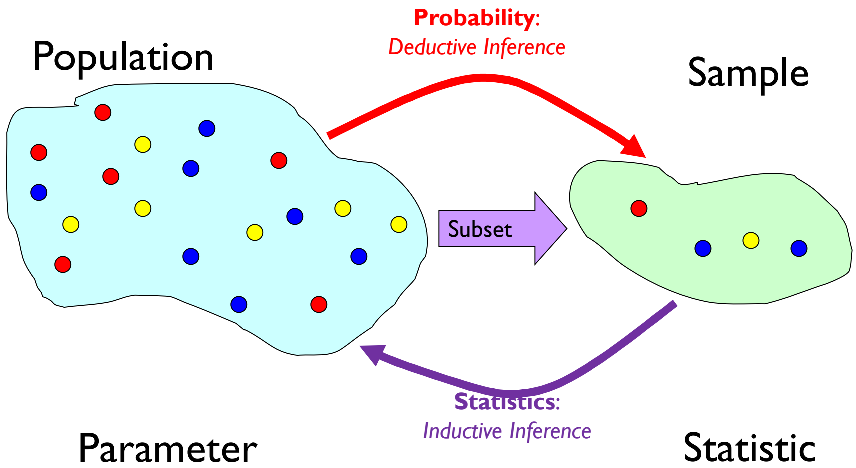
5

Probability vs Statistics



6

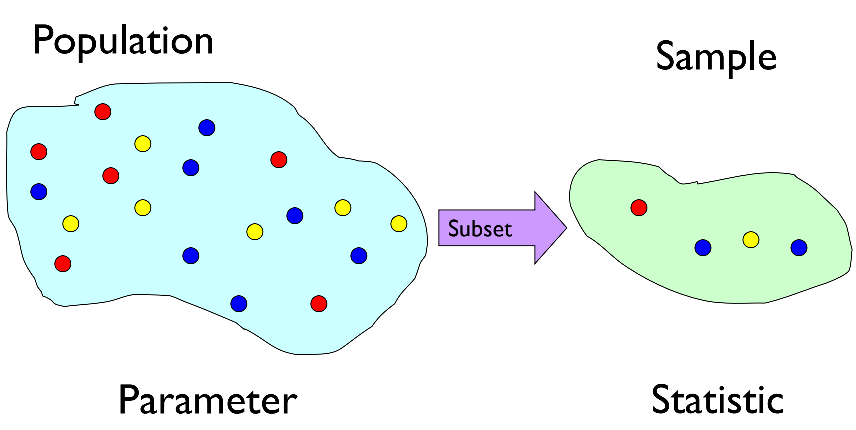
Probability vs Statistics



- Populations have Parameters,
- Samples have Statistics,

7

Key Statistical Concepts



- Populations have Parameters,
- Samples have Statistics,

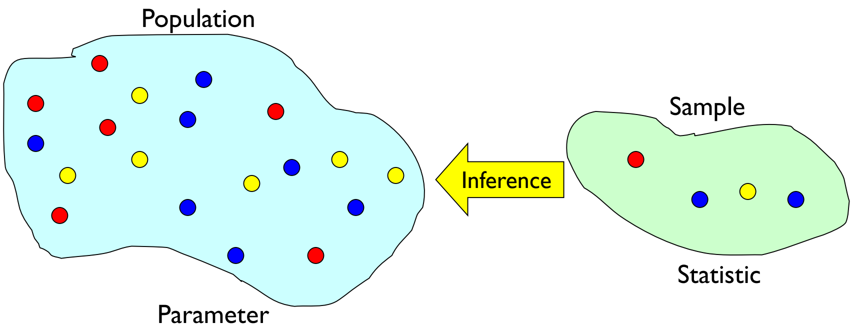
8

Descriptive Statistics

- …are ***methods*** of organizing, summarizing, and presenting data in a convenient and informative way. These methods include:
  - Graphical Techniques, and
  - Numerical Techniques.
- The actual method used depends on what ***information*** we would like to extract. Are we interested in…
  - measure(s) of central location? and/or
  - measure(s) of variability (dispersion)?
- Descriptive Statistics helps to answer these questions…

Statistical Inference

- **Statistical inference** is the ***process*** of making an estimate, prediction, or decision about a population based on a sample.



What can we ***infer*** about a Population's Parameters based on a Sample's Statistics?

Sampling Distributions

- The basic thrust of inferential statistics is drawing conclusions regarding the levels of population parameters, such as μ and σ. These conclusions can be based directly on the values of the counterpart sample statistics $\bar{X}$, s.

$\left.\begin{array}{l}\mu : \text{population mean} \\ \sigma : \text{population standard deviation}\end{array}\right\}$ Parameter

$\left.\begin{array}{l}\bar{X} \text{ or } m\text{: sample mean} \\ s : \text{sample standard deviation}\end{array}\right\}$ Statistic

Sampling Distributions

- Before the sample data are in hand, $\bar{X}$, s and other sample statistics are uncertain quantities, i.e., each is a random variable having its *probability distribution*.
- As a special class, the probability distributions for sample statistics are referred to as *sampling distributions*.
- We start with the sampling distribution for the sample mean $\bar{X}$.

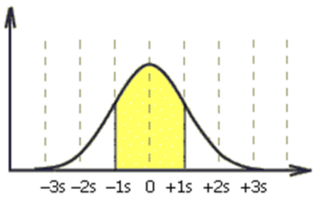
Sampling Distributions

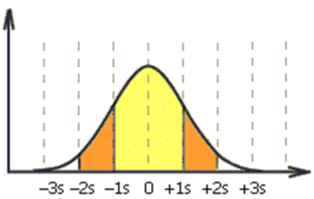
	Population	Sample
Size	N	n
Mean	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
Variance	$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}$

Note: the denominator is sample size (n) minus one !

The Empirical Rule … If the histogram is **bell shaped**

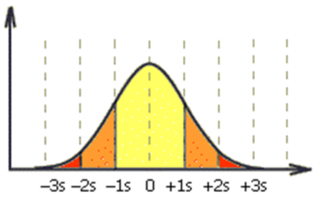
- Approximately 68% of all observations fall within one ***s*** of the *mean*.





- Approximately 95% of all observations fall within **2*s*** of the *mean*.

- Approximately 99.7% of all observations fall within **3*s*** of the *mean*.



Galton Board

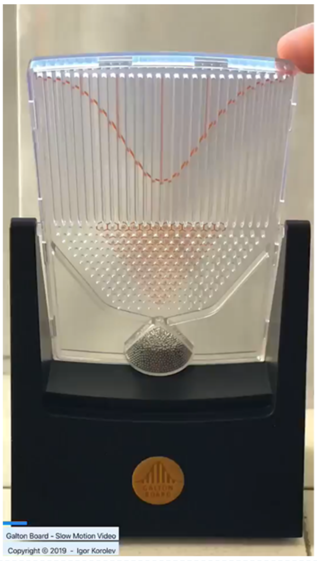
❖ ۳۰۰۰ گلوله استیل از میان ۱۲ مسیر تعبیه شده سقوط می‌کنند و شکل منحنی نرمال یا زنگوله‌ای را تشکیل می‌دهند!

❖

❖ احتمال عبور از هرکدام از این ۱۲ مسیر برای تمام گلوله‌ها 0.05 است؛ به طوری که گلوله‌ها مطابق با توزیع دو جمله‌ای، پخش می‌شوند.

❖ یکی از یافته‌های جذاب سال ۲۰۱۸، نسخه مدرن جعبه گالتون است. نسخه اصلی این جعبه، توسط فرانسیس گالتون در سال ۱۸۴۹، برای شرح تئوری حد مرکزی اختراع شد. این تئوری نشان می‌دهد که فرآیندهای تصادفی چگونه در اطراف میانگین توزیع می‌شوند.

❖ تعداد گلوله‌های ساکن در هر لوله را می‌توان با مثلث (خیام-پاسکال) قابل پیش‌بینی کرد!



15

Galton Board

16

Chebysheff's Theorem

Theorem

- Consider random variable X with mean μ and standard deviation σ, then for any $k > 0$,

$$p(|X-\mu| < k\sigma) \geq 1-1/k^2$$

Chebysheff's Theorem

- Not often used because interval is very wide.

- A more general interpretation of the standard deviation is derived from ***Chebysheff's Theorem***, which applies to all shapes of histograms (not just bell shaped).

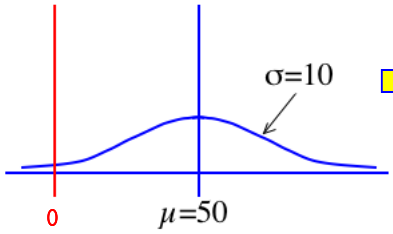
- The proportion of observations in any sample that lie within k standard deviations of the mean is *at least:*

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

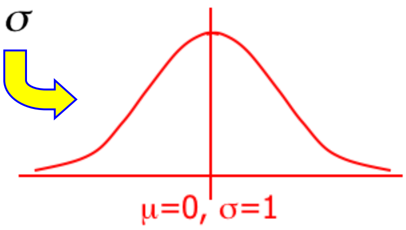
For k=2 (say), the theorem states that *at least* 3/4 of all observations lie within 2 standard deviations of the mean. This is a "lower bound" compared to Empirical Rule's approximation (95%).

Calculating Normal Probabilities

- **P(45 < X < 60)** ?



$$Z = \frac{X-\mu}{\sigma}$$



$$P(45 < X < 60) =$$

$$P\left(\frac{45-50}{10} < \frac{X-\mu}{\sigma} < \frac{60-50}{10}\right) =$$

$$P(-.5 < Z < 1)$$

19

Calculating Normal Probabilities

- How to use Normal Table

This table gives probabilities **P(0 < Z < z)**
First column = integer + first decimal
Top row = second decimal place

P(0 < Z < 0.5)

P(0 < Z < 1)

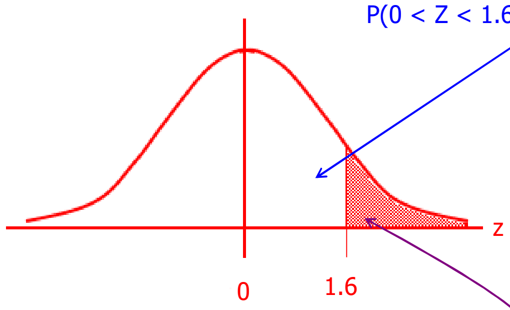
P(−0.5 < Z < 1) = .1915 + .3414 = .5328

z	.00	.01	.02	.03
0.0	.0000	.0040	.0080	.0120
0.1	.0398	.0438	.0478	.0517
0.2	.0793	.0832	.0871	.0910
0.3	.1179	.1217	.1255	.1293
0.4	.1554	.1591	.1628	.1664
0.5	.1915	.1950	.1985	.2019
0.6	.2257	.2291	.2324	.2357
0.7	.2580	.2611	.2642	.2673
0.8	.2881	.2910	.2939	.2967
0.9	.3159	.3186	.3212	.3238
1.0	.3413	.3438	.3461	.3485
1.1	.3643	.3665	.3686	.3708
1.2	[illegible]	[illegible]	[illegible]	[illegible]

20

Using the Normal Table

- What is **P(Z > 1.6)** ?



P(Z > 1.6) = .5 – P(0 < Z < 1.6)
= .5 – .4452
= **.0548**
⇒ $z_{5.48}$=1.6 (5.48%= $\alpha/2$)

21

Using the values of Z

- Because $z_{.025}$ = 1.96 and $-z_{.025}$= -1.96,

$$p(-1.96 < Z < 1.96) = 0.95$$

The interval [-1.96, 1.96] is called the *Confidence Interval of 95%* (=1-2×2.5 =1- α), and 1- α is called Confidence level.

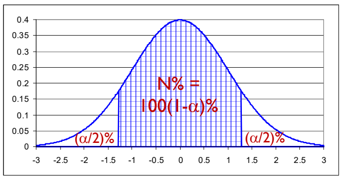
- Similarly

$$p(-1.645 < Z < 1.645) = 0.90$$

- Other Z values are

$z_{.05}$ = 1.645

$z_{.01}$ = 2.33



Confidence level N%	50%	68%	80%	90%	95%	98%	99%
Constant z_N	0.67	1.00	1.28	1.64	1.96	2.33	2.58

22

Student *t* Distribution

- Here the letter ***t*** is used to represent the random variable, hence the name. The density function for the Student ***t*** distribution is as follows…

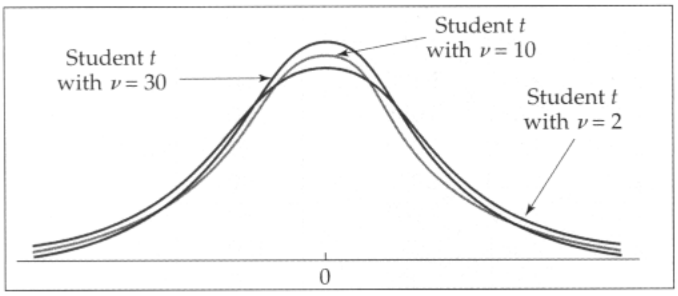
$$f(t) = \frac{\Gamma[(\nu+1)/2]}{\sqrt{\nu\pi}\,\Gamma(\nu/2)}\left[1+\frac{t^2}{\nu}\right]^{-(\nu+1)/2}$$

ν: (nu) is called ***the degrees of freedom***, i.e. (n-1) ,

Γ: (Gamma function) is $\Gamma(k) = (k-1)! = (k-1)(k-2)\ldots(2)(1)$

Student *t* Distribution…

- In much the same way that μ and σ define the normal distribution, ν , the degrees of freedom, defines the Student ***t*** Distribution:



- As the number of degrees of freedom increases, the ***t*** distribution approaches the standard normal distribution.

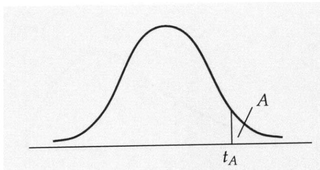
Determining Student *t* Values…

- The student **t** distribution is used extensively in statistical inference. $t_{A,\nu}$

- That is, values of a Student **t** random variable with ν degrees of freedom such that:

$$P(t > t_{A,\nu}) = A$$

- The values for A are pre-determined "critical" values, typically in the 10%, 5%, 2.5%, 1% and 1/2% range.



25

Using the *t* table

- Ex: if we want the value of **t** with 10 degrees of freedom such that the area under the Student **t** curve is .05:

Area under the curve value (t_A) : COLUMN

$t_{.05,10}$

$t_{.05,10}$=1.812

Degrees of Freedom : ROW

DEGREES OF FREEDOM	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106

26

The Sampling Distribution of the Mean

Example:

- Consider the following population: (A=4, B=3, C=2)

Name	Examination Grade	Grade Points
Ali	B	3
Kayvan	C	2
Nahid	B	3
Reza	A	4
Zohreh	C	2

$\mu = 2.8 \qquad \sigma = (0.56)^{1/2} = 0.7483$

A simple random sample of n = 2 student grade points are selected. Determine the sampling distribution of the sample mean grade points.

suppose: selection is done without replacement.

The Sampling Distribution of the Mean

Solution:

Name	Examination Grade	Grade Points
Ali	B	3
Kayvan	C	2
Nahid	B	3
Reza	A	4
Zohreh	C	2

- The probability distribution for the random variable $\bar{X}$:

Possible Mean $\bar{x}$	Student Combinations	$p(\bar{X} = \bar{x})$
2.0	(Kayvan, Zohreh)	0.1
2.5	(Ali, Kayvan) (Ali, Zohreh) (Kayvan, Nahid) (Nahid, Zohreh)	0.4
3.0	(Ali, Nahid) (Kayvan, Reza) (Reza, Zohreh)	0.3
3.5	(Ali, Reza) (Nahid, Reza)	0.2

- There are C_2^5=10 equally likely combinations of student pairs, each having probability 0.1, it follows that, for example:

$$p(\bar{X} = 3.5) = 0.2$$

- The expected value of the random variable $\bar{X}$:

$$E[\bar{X}] = \sum_{\bar{x}} \bar{x} \cdot p(\bar{X} = \bar{x}) = (2\times 0.1) + (2.5\times 0.4) + (3\times 0.3) + (3.5\times 0.2) = 2.8$$

The Sampling Distribution of the Mean

- The mean of the population or μ:

$$\mu = 2.8$$

Note: Property of the sample mean:

$$E[\overline{X}] = \mu$$

- The standard deviation of $\overline{X}$, i.e. $\sigma_{\overline{X}}$:

$$E[\overline{X}^2] = \sum_{\overline{x}} \overline{x}^2 \cdot p(\overline{X} = \overline{x}) = (2^2 \times 0.1) + (2.5^2 \times 0.4) + (3^2 \times 0.3) + (3.5^2 \times 0.2) = 8.05$$

$$Var(\overline{X}) = E[\overline{X}^2] - E[\overline{X}]^2 = 8.05 - 2.8^2 = 0.21$$

$$\sigma_{\overline{x}} = SD(\overline{X}) = \sqrt{0.21} = 0.458$$

- The population standard deviation:

$$\sigma = 0.7483$$

The Sampling Distribution of the Mean

- For small population:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

- For large population

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

The Sampling Distribution of the Mean

Example:

- A die is thrown infinitely many times. Let X represent the number of spots showing on any throw.
- The probability distribution of X is

x	1	2	3	4	5	6
$p(x)$	1/6	1/6	1/6	1/6	1/6	1/6

$E(X) = 1(1/6) + 2(1/6) + 3(1/6) + \ldots = 3.5$
$V(X) = (1-3.5)^2(1/6) + (2-3.5)^2(1/6) + \ldots = 2.92$

The Sampling Distribution of the Mean

Example :

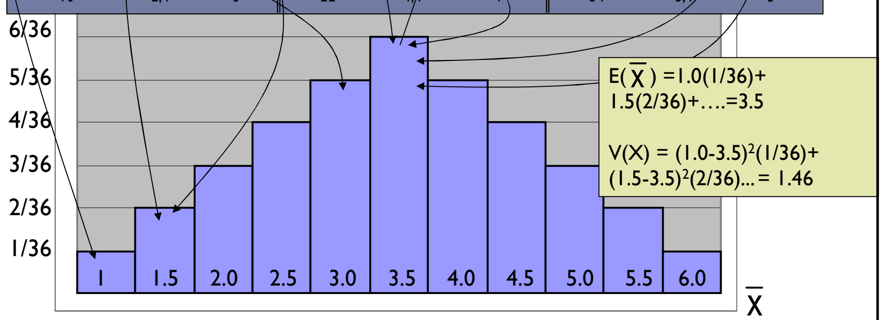
Throwing a die twice:

- Suppose we want to estimate μ from the mean $\bar{X}$ of a sample of size n = 2.
- What is the distribution of $\bar{X}$?

The Sampling Distribution of the Mean

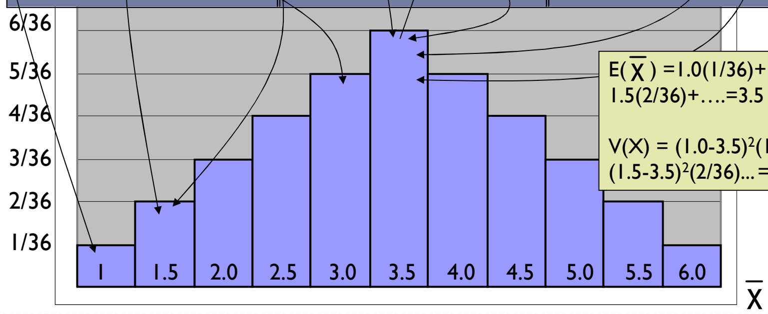
Sample		Mean	Sample		Mean	Sample		Mean
1	1,1	1	13	3,1	2	25	5,1	3
2	1,2	1.5	14	3,2	2.5	26	5,2	3.5
3	1,3	2	15	3,3	3	27	5,3	4
4	1,4	2.5	16	3,4	3.5	28	5,4	4.5
5	1,5	3	17	3,5	4	29	5,5	5
6	1,6	3.5	18	3,6	4.5	30	5,6	5.5
7	2,1	1.5	19	4,1	2.5	31	6,1	3.5
8	2,2	2	20	4,2	3	32	6,2	4
9	2,3	2.5	21	4,3	3.5	33	6,3	4.5
10	2,4	3	22	4,4	4	34	6,4	5
11	2,5	3.5	23	4,5	4.5	35	6,5	5.5
12	2,6	4	24	4,6	5	36	6,6	6

Sample		Mean	Sample		Mean	Sample		Mean
1	1,1	1	13	3,1	2	25	5,1	3
2	1,2	1.5	14	3,2	2.5	26	5,2	3.5
3	1,3	2	15	3,3	3	27	5,3	4
4	1,4	2.5	16	3,4	3.5	28	5,4	4.5
5	1,5	3	17	3,5	4	29	5,5	5
6	1,6	3.5	18	3,6	4.5	30	5,6	5.5
7	2,1	1.5	19	4,1	2.5	31	6,1	3.5
8	2,2	2	20	4,2	3	32	6,2	4
9	2,3	2.5	21	4,3	3.5	33	6,3	4.5
10	2,4	3	22	4,4	4	34	6,4	5



Sample		Mean	Sample		Mean	Sample		Mean
1	1,1	1	13	3,1	2	25	5,1	3
2	1,2	1.5	14	3,2	2.5	26	5,2	3.5
3	1,3	2	15	3,3	3	27	5,3	4
4	1,4	2.5	16	3,4	3.5	28	5,4	4.5
5	1							5
6	1							5.5
7	2							3.5
8	2							4
9	2,3	2.5	21	4,3	3.5	33	6,3	4.5
10	2,4	3	22	4,4	4	34	6,4	5

$$Note: \mu_{\overline{x}} = \mu_x \quad and \quad \sigma_{\overline{x}}^{\;2} = \frac{\sigma^2}{2}$$



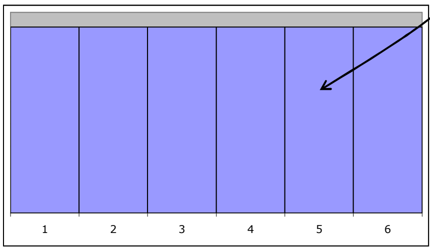
$E(\overline{X}) = 1.0(1/36)+1.5(2/36)+\ldots=3.5$

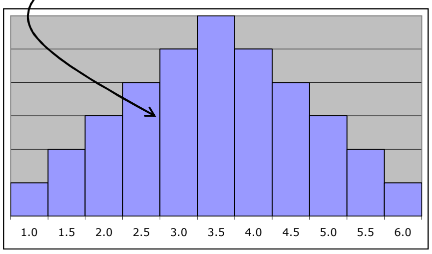
$V(X) = (1.0-3.5)^2(1/36)+(1.5-3.5)^2(2/36)\ldots= 1.46$

35

The Sampling Distribution of the Mean

- Compare the distribution of X and $\overline{X}$





- As well, note that:

$$\mu_{\overline{x}} = \mu$$
$$\sigma_{\overline{x}}^2 = \sigma^2/2$$

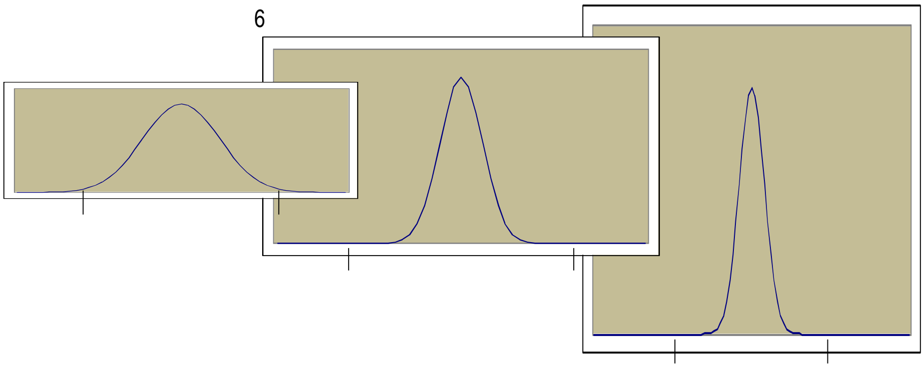
36

The Sampling Distribution of the Mean

$n = 5$
$\mu_{\bar{x}} = 3.5$
$\sigma^2_{\bar{x}} = .5833\ (= \frac{\sigma^2_x}{5})$

$n = 10$
$\mu_{\bar{x}} = 3.5$
$\sigma^2_{\bar{x}} = .2917\ (= \frac{\sigma^2_x}{10})$

$n = 25$
$\mu_{\bar{x}} = 3.5$
$\sigma^2_{\bar{x}} = .1167\ (= \frac{\sigma^2_x}{25})$



37

The Sampling Distribution of the Mean

$n = 5$
$\mu_{\bar{x}} = 3.5$
$\sigma^2_{\bar{x}} = .5833\ (= \frac{\sigma^2_x}{5})$

$n = 10$
$\mu_{\bar{x}} = 3.5$
$\sigma^2_{\bar{x}} = .2917\ (= \frac{\sigma^2_x}{10})$

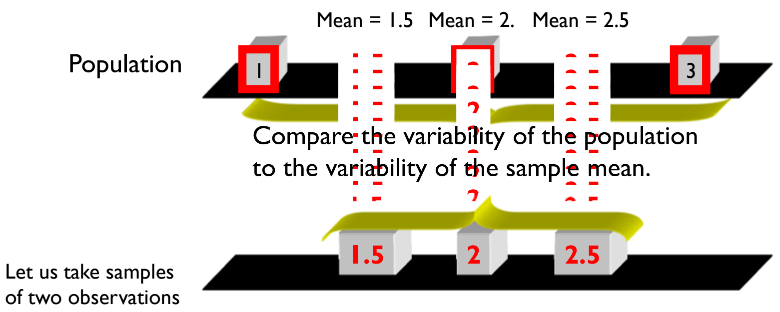
$n = 25$
$\mu_{\bar{x}} = 3.5$
$\sigma^2_{\bar{x}} = .1167\ (= \frac{\sigma^2_x}{25})$

Notice that $\sigma^2_{\bar{x}}$ is smaller than σ^2_x. The larger the sample size the smaller $\sigma^2_{\bar{x}}$. Therefore, $\bar{X}$ tends to fall closer to μ, as the sample size increases.

38

The Sampling Distribution of the Mean

Demonstration: The variance of the sample mean is smaller than the variance of the population.



The Sampling Distribution of the Mean

Also,
Expected value of the population =
(1 + 2 + 3)/3 = 2

Expected value of the sample mean =
(1.5 + 2 + 2.5)/3 = 2

The Sampling Distribution of the Mean

- Probabilities for the random variable $\overline{X}$ may be obtained from the standard normal curve:

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

- In most applications, σ is unknown, thus sample standard deviation (s) is used:

$$\frac{\bar{X} - \mu}{s/\sqrt{n}}$$

- Before data are collected, $\overline{X}$ and s are random variables, thus above expression would be a composite random variable that is itself a function of $\overline{X}$ and s.

The Sampling Distribution of the Mean

- W.S. Gosset published his results about the composite random variable discussed in the previous slide, and in his honor, it is mentioned as the *Student t statistics*:

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

- The above composite random variable is a continuous random variable whose probability distribution is ***Student t***, and is completely specified by a single parameter, referred to as the number of degrees of freedom (df):

df=*n*-1

The Sampling Distribution of the Mean

By the Central Limit Theorem:

- If a random sample is drawn from any population, the sampling distribution of the sample mean is approximately normal for a sufficiently large sample size.
- The larger the sample size, the more closely the sampling distribution of $\overline{X}$ will resemble a normal distribution.

The Sampling Distribution of the Mean

1. $\mu_{\bar{x}} = \mu_x$

2. $\sigma_{\bar{x}}^2 = \dfrac{\sigma_x^2}{n}$

3. If x is ***normal***, $\overline{X}$ is ***normal.***

   If x is ***nonnormal*** $\overline{X}$ is ***approximately normally distributed*** for sufficiently large sample size.

Estimation…

- There are two types of inference:
  - Estimation,
  - Hypothesis testing;

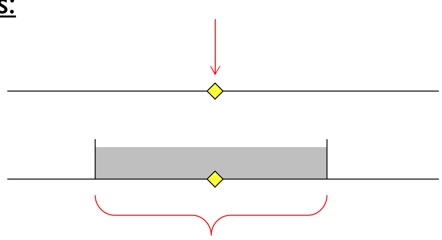
- ***estimation*** will be talked.

45

Estimation…

- The objective of estimation is to determine the ***approximate value*** of a population parameter on the basis of a sample statistic.
  - E.g., the sample mean ($\overline{x}$) is employed to ***estimate*** the population mean (μ).

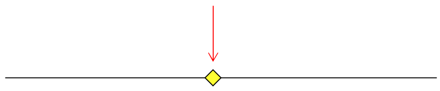
- <u>There are two types of estimators:</u>
  - *Point Estimator*
  - *Interval Estimator*



46

Point Estimator…

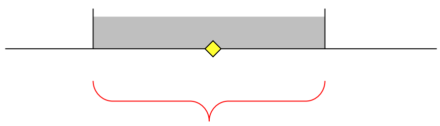
- A ***point estimator*** draws inferences about a population by estimating the value of an unknown parameter using a single value or point.



- Point probabilities in continuous distributions were virtually zero. Likewise, we'd expect that the point estimator gets closer to the parameter value with an increased sample size but point estimators don't reflect the effects of larger sample sizes. Hence, we will employ the ***interval estimator*** to estimate population parameters…

Interval Estimator

- An ***interval estimator*** draws inferences about a population by estimating the value of an unknown parameter using an interval.



- That is we say (with some N% certainty) that the population parameter of interest is between some lower and upper bounds.

Point & Interval Estimation…

- For example, suppose we want to estimate the mean score of a class of ML students. For n=25 students,
- $\overline{X}$ is calculated to be 14.5.

point estimate

interval estimate

- An alternative statement is:

  The mean income is ***between*** 12.5 and 15.5.

Qualities of Estimators

- Qualities desirable in estimators include unbiasedness, consistency, and relative efficiency:
  - An ***unbiased estimator*** of a population parameter is an estimator whose expected value is equal to that parameter.

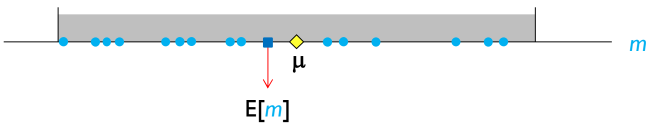
    $$? = E[Y] - p$$

  - An ***unbiased estimator*** is said to be ***consistent*** if the difference between the estimator and the parameter becomes smaller as the sample size (n) grows larger, i.e., $n \to \infty$. In other words, the variance of the estimator should vanish.
  - If there are two ***unbiased estimators*** of a parameter, the one whose variance is smaller is said to be ***relatively efficient***.

Example

Example 1: (estimator of the mean)

- Suppose sample $\mathcal{X} = \{x^t\}_{t=1}^{n}$ drawn from some population with some density with mean μ. Show that the sample average, m, is an unbiased & consistent estimator of the mean, μ,
  - Solution:



$$E[m] = E\left[\frac{\sum_t x^t}{n}\right] = \frac{1}{n}\sum_t E[x^t] = \frac{n\,\mu}{n} = \mu$$

- This means that though on a particular sample, m may be different from μ, if we take many such samples, $\mathcal{X}_i$, and estimate many $m_i = m(\mathcal{X}_i)$, their average will get close to μ as the number of such samples increases.

Example

- m is also a consistent estimator, i.e.,

$$\mathrm{Var}(m) \to 0 \text{ as } n \to \infty$$

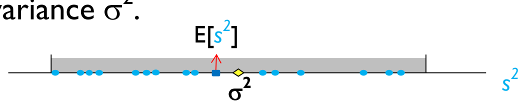
because:

$$\mathrm{Var}(m) = \mathrm{Var}\left(\frac{\sum_t x^t}{n}\right) = \frac{1}{n^2}\sum_t \mathrm{Var}(x^t) = \frac{n\,\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Example

Example 2: (estimator of the variance)

- Suppose sample $\mathcal{X} = \{ x^t \}_{t=1}^{n}$ drawn from some population with some density with variance σ^2. Show that the sample variance, s^2, is a biased estimator of the variance σ^2.
  - Solution:



- the MLE of σ^2:

$$s^2 = \frac{\sum_t (x^t - m)^2}{n} = \frac{\sum_t (x^t)^2 - n\, m^2}{n}$$

$$E[s^2] = \frac{\sum_t E[(x^t)^2] - n \cdot E[m^2]}{n} \qquad \text{1}$$

- Note that $\mathrm{Var}(Z) = E[Z^2] - E[Z]^2$, and we get $E[Z^2] = \mathrm{Var}(Z) + E[Z]^2$
- Thus, we can write:

$$E[(x^t)^2] = \sigma^2 + \mu^2 \qquad \text{2}$$

- From the example1, we have know $E[m]$ and $\mathrm{Vra(m)}$, thus:

$$E[m^2] = \mathrm{Var(m)} + E[m]^2 = \sigma^2/n + \mu^2 \qquad \text{3}$$

53

Example

$$\left.\begin{array}{c} 1 \\ 2 \\ 3 \end{array}\right] \Rightarrow E[s^2] = \frac{n(\sigma^2 + \mu^2) - n(\sigma^2/n + \mu^2)}{n} = \left(\frac{n-1}{n}\right)\sigma^2 \neq \sigma^2$$

- which shows that s^2 is a biased estimator of σ^2. $(n/(n-1))\, s^2$ is an unbiased estimator. However when sample size (n) is large, the difference is negligible. This is an example of an asymptotically unbiased estimator whose bias goes to 0 as n goes to infinity.

- Question: Is s^2 a consistent estimator of σ^2 ?

54

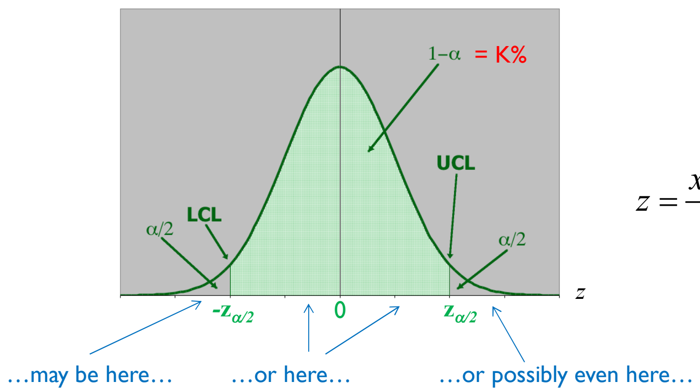
Sampling Distributions

	Population	Sample
Size	Pop	n
Mean	$\mu = \dfrac{\sum_{i=1}^{Pop} x_i}{Pop}$	$\bar{X} = \dfrac{\sum_{i=1}^{n} x_i}{n}$
Variance	$\sigma^2 = \dfrac{\sum_{i=1}^{Pop}(x_i - \mu)^2}{Pop}$	$s^2 = \dfrac{\sum_{i=1}^{N}(x_i - \bar{X})^2}{n-1}$

Note: the denominator is sample size (n) minus one !

Confidence Interval

- The probability $1-\alpha$ is called the ***confidence level***.



$$z = \frac{x - \mu}{\sigma}$$

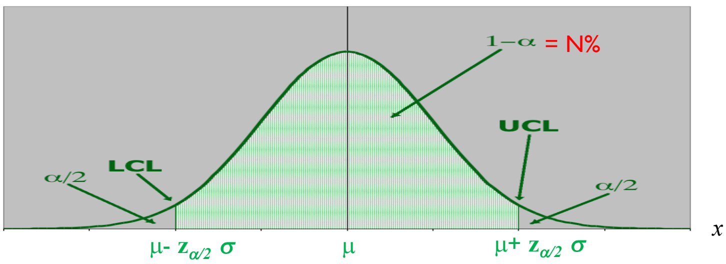
Four commonly used confidence levels…

- Confidence Level

$1-\alpha$	α	$\alpha/2$	$z_{\alpha/2}$
.90	.10	.05	$z_{.05} = 1.645$
.95	.05	.025	$z_{.025} = 1.96$
.98	.02	.01	$z_{.01} = 2.33$
.99	.01	.005	$z_{.005} = 2.575$

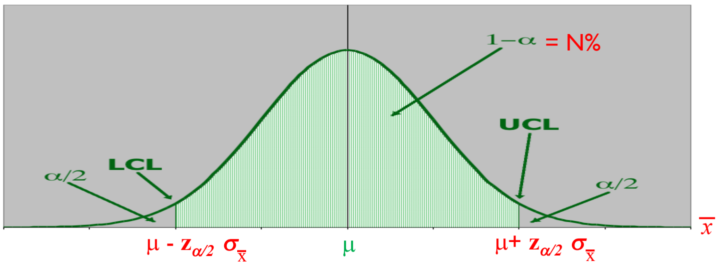
Confidence Interval

- The probability $1-\alpha$ is called the ***confidence level***.



Confidence Interval Estimator for μ

- $\bar{X}$ or m is the estimator for μ, thus…



Note:

The population mean is a fixed but ***unknown*** quantity. Its incorrect to interpret the confidence interval estimate as a probability statement about μ. The interval acts as the lower and upper limits of the interval ***estimate*** of the population mean.

Confidence Interval Estimator for μ

- As said before, to make inference about population parameters we use sampling distributions. … the ***actual*** location of the population mean μ is:

$$\mu - z_n\sigma_{\bar{X}} \le \bar{X} \le \mu + z_n\sigma_{\bar{X}} \quad \Rightarrow \quad \bar{X} - z_n\sigma_{\bar{X}} \le \mu \le \bar{X} + z_n\sigma_{\bar{X}}$$

Usually represented with a "plus/minus" (±) sign

upper confidence limit (UCL)

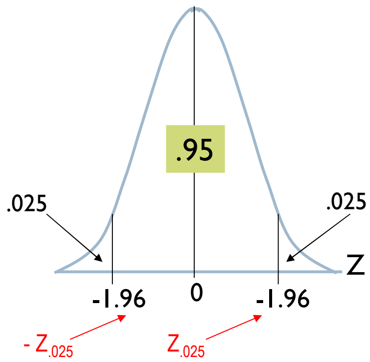
$$\bar{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = \left\{ \bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right\}$$

lower confidence limit (LCL)

Confidence Interval Estimator for μ

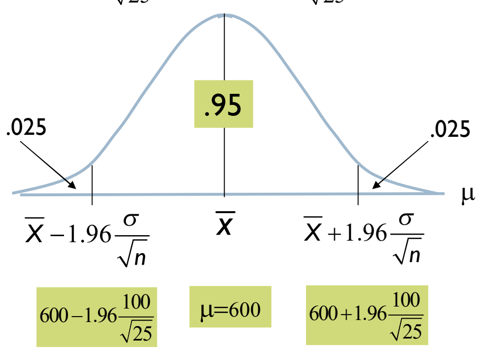
Example: $\overline{X} = 600,\ n=25,\ \sigma=100$

Standard normal distribution Z



Normal distribution of $\overline{X}$

$$p(600-1.96\frac{100}{\sqrt{25}} \le \overline{x} \le 600+1.96\frac{100}{\sqrt{25}}) = .95$$



Confidence Interval Estimator for μ

Example: $= 600,\ n=25,\ \sigma=100$

$$p(600-1.96\frac{100}{\sqrt{25}} \le \mu \le 600+1.96\frac{100}{\sqrt{25}}) = .95$$

Which reduces to

$$p(560.8 \le \mu \le 639.2) = .95$$

- Conclusion
  - There is 95% chance that the population mean falls within the interval [560.8, 639.2] **if the sample mean is 600.**

❖ **Note:**

- In fact, we don't know the standard deviation of the population!!

Interval Width…

- The width of the confidence interval estimate is a function of:
  - the confidence level,
  - the population standard deviation,
  - and the sample size…

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

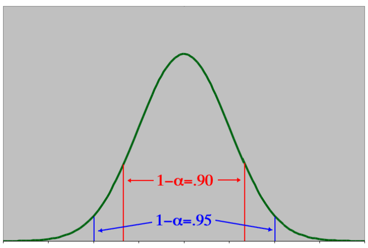
63

Interval Width…

- The width of the confidence interval estimate is a function of the confidence level, the population standard deviation, and the sample size…

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- A larger confidence level produces a **w i d e r** confidence interval



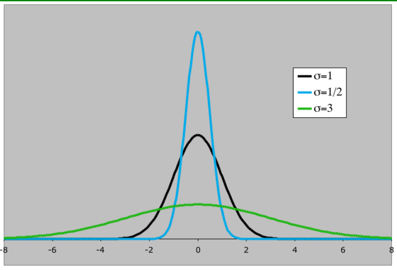
64

Interval Width…

- The width of the confidence interval estimate is a function of the confidence level, the population standard deviation, and the sample size…

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- Larger values of σ produce **w i d e r** confidence intervals



Interval Width…

- The width of the confidence interval estimate is a function of the confidence level, the population standard deviation, and the sample size…

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- Increasing the sample size decreases the width of the confidence interval while the confidence level can remain unchanged.

- Note: this also increases the ***cost*** of obtaining additional data

Sampling Distribution: Difference of two means

- The ***expected value*** and ***variance*** of the sampling distribution of are given by:

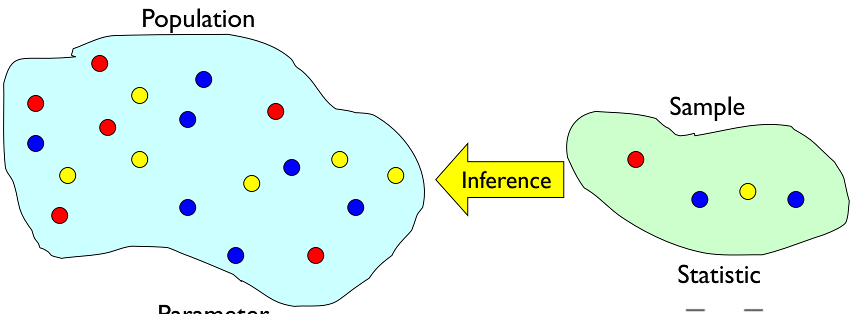
$$\bar{X}_1 - \bar{X}_2$$

- mean: $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$

- standard deviation: $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$

(also called the standard error if the difference between two means)

Sampling Distribution: Difference of two means



$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$$

$$\bar{X}_1 - \bar{X}_2$$