Computing Confidence Intervals for Sample Data

Topics

- Use of Statistics
- Sources of errors
- Accuracy, precision, resolution
- A mathematical model of errors
- Confidence intervals
  - For means
  - For variances
  - For proportions
- How many measurements are needed for desired error?

What are *statistics*?

- "A branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data."

  Merriam-Webster

→ We are most interested in analysis and interpretation here.

- "Lies, damn lies, and statistics!"

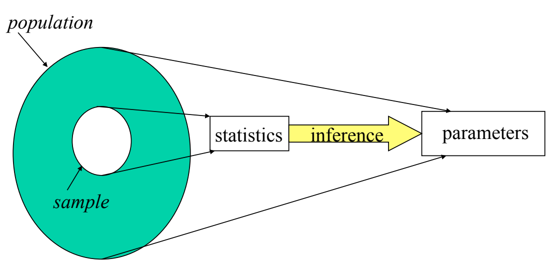
What is a *statistic*?

- "A quantity that is computed from a sample [of data]."

  Merriam-Webster

- An estimate of a population parameter

Statistical Inference



5

Why do we need statistics?

- A set of experimental measurements constitute a sample of the underlying process/system being measured
  - Use statistical techniques to infer the true value of the metric
- Use statistical techniques to quantify the amount of imprecision due to random experimental errors

6

Experimental errors

- Errors → *noise* in measured values
- *Systematic* errors
  - Result of an experimental "mistake"
  - Typically produce constant or slowly varying bias
- Controlled through skill of experimenter
- Examples
  - Temperature change causes clock drift
  - Forget to clear cache before timing run

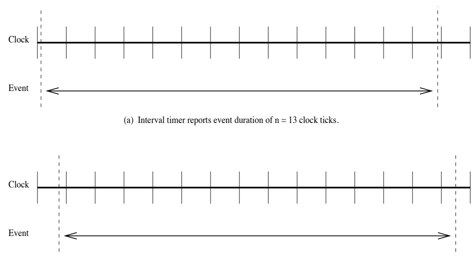
7

Experimental errors

- *Random* errors
  - Unpredictable, non-deterministic
  - Unbiased → equal probability of increasing or decreasing measured value
- Result of
  - Limitations of measuring tool
  - Observer reading output of tool
  - Random processes within system
- Typically cannot be controlled
  - Use statistical tools to characterize and quantify

8

Example: Quantization → Random error



(a) Interval timer reports event duration of n = 13 clock ticks.

(b) Interval timer reports event duration of n = 14 clock ticks.

9

Quantization error

- Timer resolution
  → quantization error
- Repeated measurements
  X ± Δ
  Completely unpredictable

10

A Model of Errors

Error	*Measured value*	*Probability*
-E	*x* – E	½
+E	*x* + E	½

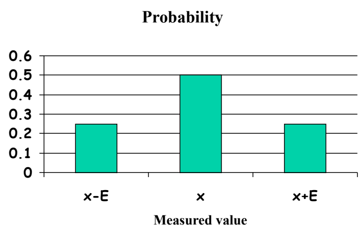
11

A Model of Errors

Error 1	*Error 2*	*Measured value*	*Probability*
-E	-E	*x* – 2E	¼
-E	+E	*x*	¼
+E	-E	*x*	¼
+E	+E	*x* + 2E	¼

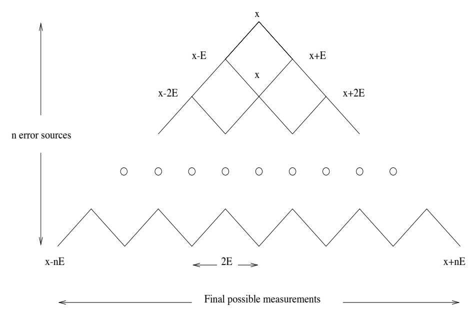
12

A Model of Errors



13

Probability of Obtaining a Specific Measured Value



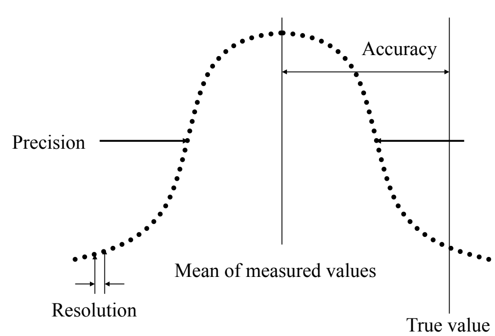
14

A Model of Errors

- Pr(X=x_i) = Pr(measure x_i)
  = number of paths from real value to x_i
- Pr(X=x_i) ~ binomial distribution
- As number of error sources becomes large
  - $n \rightarrow \infty$,
  - Binomial → Gaussian (Normal)
- Thus, the bell curve

15

Frequency of Measuring Specific Values



16

Accuracy, Precision, Resolution

- Systematic errors → accuracy
  - How close mean of measured values is to true value
- Random errors → precision
  - Repeatability of measurements
- Characteristics of tools → resolution
  - Smallest increment between measured values

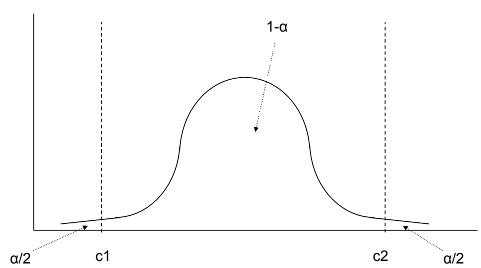
17

Quantifying Accuracy, Precision, Resolution

- Accuracy
  - Hard to determine true accuracy
  - Relative to a predefined standard
    - E.g. definition of a "second"
- Resolution
  - Dependent on tools
- Precision
  - Quantify amount of *imprecision* using statistical tools

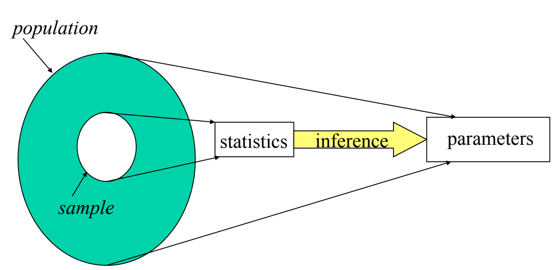
18

Confidence Interval for the Mean



19

Statistical Inference

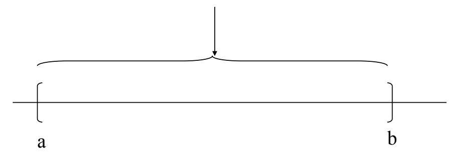


20

Why do we need statistics?

- A set of experimental measurements constitute a sample of the underlying process/system being measured
  - Use statistical techniques to infer the true value of the metric
- Use statistical techniques to quantify the amount of imprecision due to random experimental errors
  - Assumption: random errors normally distributed

Interval Estimate



The interval estimate of the population parameter will have a specified confidence or probability of correctly estimating the population parameter.

Properties of Point Estimators

- In statistics, **point estimation** involves the use of sample data to calculate a single value which is to serve as a "best guess" for an unknown (fixed or random) population parameter.
- Example of point estimator: sample mean.
- Properties:
  - Unbiasedness: the expected value of all possible sample statistics (of given size *n*) is equal to the population parameter.
    $$E[\overline{X}] = \mu$$
    $$E[s^2] = \sigma^2$$
  - Efficiency: precision as estimator of the population parameter.
  - Consistency: as the sample size increases the sample statistic becomes a better estimator of the population parameter.

Unbiasedness of the Mean

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$E[\overline{X}] = \frac{E\left[\sum_{i=1}^{n} X_i\right]}{n} = \frac{\sum_{i=1}^{n} E[X_i]}{n} =$$

$$\frac{\sum_{i=1}^{n} \mu}{n} = \frac{n\mu}{n} = \mu$$

Sample size= 15 — ***1.7% of population***

	Sample 1	Sample 2	Sample 3	E[sample]	Population	Error
	0.0739	0.0202	0.2918			
	0.1407	0.1089	0.4696			
	0.1257	0.0242	0.8644			
	0.0432	0.4253	0.1494			
	0.1784	0.1584	0.4242			
	0.4106	0.8948	0.0051			
	0.1514	0.0352	1.1706			
	0.4542	0.1752	0.0084			
	0.0485	0.3287	0.0600			
	0.1705	0.1697	0.7820			
	0.3335	0.0920	0.4985			
	0.1772	0.1488	0.0988			
	0.0242	0.2486	0.4896			
	0.2183	0.4627	0.1892			
	0.0274	0.4079	0.1142			
Sample Average	0.1718	0.2467	0.3744	0.2643	0.2083	26.9%
Sample Variance	0.0180	0.0534	0.1204	0.0639	0.0440	45.3%
Efficiency (average)	18%	18%	80%			
Efficiency (variance)	59%	21%	173%			

25

Sample size = 87 — ***10% of population***

	Sample 1	Sample 2	Sample 3		Population	% Rel. Error
	0.5725	0.3864	0.4627			
	0.0701	0.0488	0.2317			
	0.2165	0.0611	0.1138			
	0.6581	0.0881	0.0047			
	0.0440	0.5866	0.2438			
	0.1777	0.3419	0.0819			
	0.2380	0.1923	0.6581			
	0.0102	0.9460	0.0714			
	0.4325	0.0445	0.2959			
Sample Average	0.2239	0.2203	0.2178	0.2206	0.2083	5.9%
Sample Variance	0.0452688	0.0484057	0.0440444	0.0459	0.0440	4.3%
Efficiency (average)	7.5%	5.7%	4.5%			
Efficiency (variance)	2.9%	10.0%	0.1%			

26

Confidence Interval Estimation of the Mean

- Known population standard deviation.
- Unknown population standard deviation:
  - Large samples: sample standard deviation is a good estimate for population standard deviation. OK to use normal distribution.
  - Small samples and original variable is normally distributed: use *t* distribution with *n-1* degrees of freedom.

27

Central Limit Theorem

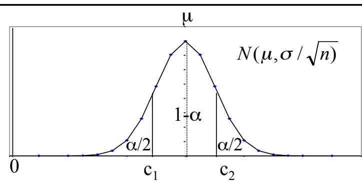
- If the observations in a sample are independent and come from the same population that has mean μ and standard deviation σ then the sample mean for **large** samples has a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$

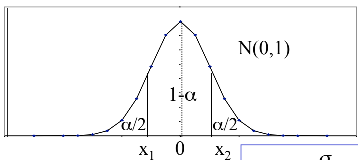
$$\bar{x} \sim N(\mu, \sigma/\sqrt{n})$$

- The standard deviation of the sample mean is called the *standard error.*

28

$c_2 = Q(1-\alpha/2)$
$c_1 = Q(\alpha/2)$





$x_2 = z_{1-\alpha/2}$
$x_1 = z_{\alpha/2} = -z_{1-\alpha/2}$

$$c_2 = \mu + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}$$

$$c_1 = \mu + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} = \mu - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}$$

29

Confidence Interval - large (n>30) samples

- 100 (1-α)% confidence interval for the population mean:

$$\left(\overline{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}}, \overline{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}}\right)$$

$\overline{x}$: sample mean
s: sample standard deviation
n: sample size
$z_{1-\alpha/2}$: (1-α/2)-quantile of a unit normal variate (N(0,1)).

30

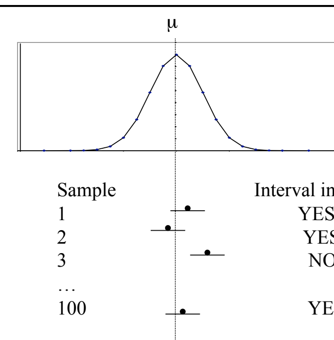
	0.4325	0.0445	0.2959		**Population**
Sample Average	0.2239	0.2203	0.2178	0.2206	**0.2083**
Sample Variance	0.0452688	0.0484057	0.0440444	0.0459	**0.0440**
Efficiency (average)	7.5%	5.7%	4.5%		
Efficiency (variance)	2.9%	10.0%	0.1%		
95% interval lower	0.1792	0.1740	0.1737		
95% interval upper	0.2686	0.2665	0.2619	0.0894	
Mean in interval	YES	YES	YES		
99% interval lower	0.1651	0.1595	0.1598		
99% interval upper	0.2826	0.2810	0.2757	0.1175	
Mean in interval	YES	YES	YES		
90% interval lower	0.1864	0.1815	0.1807		
90% interval upper	0.2614	0.2591	0.2548	0.0750	
Mean in interval	YES	YES	YES		

In Excel:
½ interval = CONFIDENCE(1-0.95,s,n)
α

interval size

Note that the higher the confidence level the larger the interval

31



Sample	Interval include μ?
1	YES
2	YES
3	NO
…	
100	YES

100 (1 – α) of the 100 samples include the population mean μ.

32

Confidence Interval Estimation of the Mean

- Known population standard deviation.
- Unknown population standard deviation:
  - Large samples: sample standard deviation is a good estimate for population standard deviation. OK to use normal distribution.
  - Small samples and original variable is normally distributed: use *t* distribution with *n-1* degrees of freedom.

33

Student's t distribution

$$t(v) \sim \frac{N(0,1)}{\sqrt{\chi^2(v)/v}}$$

v: number of degrees of freedom.

$\chi^2(v)$: chi-square distribution with *v* degrees of freedom. Equal to the sum of squares of *v* unit normal variates.

- the pdf of a t-variate is similar to that of a N(0,1).
- for *v* > 30 a t distribution can be approximated by N(0,1).

34

Confidence Interval (small samples)

- For samples from a normal distribution N(μ,σ²), $(\bar{X}-\mu)/(\sigma/\sqrt{n})$ has a N(0,1) distribution and $(n-1)s^2/\sigma^2$ has a chi-square distribution with n-1 degrees of freedom
- Thus, $(\bar{X}-\mu)/\sqrt{s^2/n}$ has a t distribution with n-1 degrees of freedom

35

Confidence Interval (small samples, normally distributed population)

- 100 (1-α)% confidence interval for the population mean:

$$(\bar{x} - t_{[1-\alpha/2;n-1]}\frac{s}{\sqrt{n}}, \bar{x} + t_{[1-\alpha/2;n-1]}\frac{s}{\sqrt{n}})$$

$\bar{x}$: sample mean
s: sample standard deviation
n: sample size
$t_{[1-\alpha/2;n-1]}$: critical value of the *t* distribution with *n-1* degrees of freedom for an area of α/2 for the upper tail.

36

Using the *t* Distribution. Sample size= 15.

	0.0274	0.4079	0.1142	E[sample]	Population	Error
Sample Average	0.1718	0.2467	0.3744	0.2643	0.2083	26.9%
Sample Variance	0.0180	0.0534	0.1204	0.0639	0.0440	45.3%
Efficiency (average)	18%	18%	80%			
Efficiency (variance)	59%	21%	173%			
95% interval lower	0.0975	0.1187	0.1823		95%,n-1 critical value	2.145
95% interval upper	0.2462	0.3747	0.5665			
Mean in inteval	YES	YES	YES			

In Excel: TINV(1-0.95,15-1)

α

How many measurements do we need for a desired interval width?

- Width of interval inversely proportional to √n
- Want to minimize number of measurements
- Find confidence interval for mean, such that:
  - Pr(actual mean in interval) = (1 – α)

$$(c_1, c_2) = \left[(1-e)\bar{x}, (1+e)\bar{x}\right]$$

How many measurements?

$$(c_1, c_2) = (1 \mp e)\bar{x}$$

$$= \bar{x} \mp z_{1-\alpha/2} \frac{s}{\sqrt{n}}$$

$$z_{1-\alpha/2} \frac{s}{\sqrt{n}} = \bar{x}e$$

$$n = \left(\frac{z_{1-\alpha/2}\, s}{\bar{x}e}\right)^2$$

How many measurements?

- But n depends on knowing mean and standard deviation!
- Estimate *s* with small number of measurements
- Use this *s* to find *n* needed for desired interval width

How many measurements?

- Mean = 7.94 s
- Standard deviation = 2.14 s
- Want 90% confidence mean is within 7% of actual mean.

41

How many measurements?

- Mean = 7.94 s
- Standard deviation = 2.14 s
- Want 90% confidence mean is within 7% of actual mean.
- $\alpha = 0.90$
- $(1-\alpha/2) = 0.95$
- Error = ± 3.5%
- $e = 0.035$

42

How many measurements?

$$n = \left(\frac{z_{1-\alpha/2}s}{\bar{x}e}\right)^2 = \left(\frac{1.895(2.14)}{0.035(7.94)}\right) = 212.9$$

- 213 measurements

→ 90% chance true mean is within ± 3.5% interval

43

Confidence Interval Estimates for Proportions

Confidence Interval for Proportions

- For categorical data:
  - E.g. file types {html, html, gif, jpg, html, pdf, ps, html, pdf …}
  - If n_1 of n observations are of type html, then the sample proportion of html files is $p = n_1/n$.
- The population proportion is π.
- Goal: provide confidence interval for the population proportion π.

Confidence Interval for Proportions

- The sampling distribution of the proportion formed by computing p from all possible samples of size n from a population of size N with replacement tends to a normal with mean π and standard error $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$.
- The normal distribution is being used to approximate the binomial. So, $n\pi \geq 10$

Confidence Interval for Proportions

- The (1-α)% confidence interval for π is

$$\left(p - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}},\ p + z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right)$$

p: sample proportion.
n: sample size
$z_{1-\alpha/2}$: (1-α/2)-quantile of a unit normal variate (N(0,1)).

Example 1

One thousand entries are selected from a Web log. Six hundred and fifty correspond to gif files. Find 90% and 95% confidence intervals for the proportion of files that are gif files.

Sample size (n)	1000		
No. gif files in sample	650		
Sample proportion (p)	0.65		
n*p	650	> 10	OK

90% confidence interval

alpha	0.1
1-alpha/2	0.95
z0.95	1.645
Lower bound	0.625
Upper bound	0.675

In Excel:
NORMSINV(1-0.1/2)

95% confidence interval

alpha	0.05
1-alpha/2	0.975
z0.975	1.960
Lower bound	0.620
Upper bound	0.680

NORMSINV(1-0.05/2)

Example 2

- How much time does processor spend in OS?
- Interrupt every 10 ms
- Increment counters
  - n = number of interrupts
  - m = number of interrupts when PC within OS

49

Proportions

- How much time does processor spend in OS?
- Interrupt every 10 ms
- Increment counters
  - n = number of interrupts
  - m = number of interrupts when PC within OS
- Run for 1 minute
  - n = 6000
  - m = 658

50

Proportions

$$(c_1, c_2) = \bar{p} \mp z_{1-\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$= 0.1097 \mp 1.96\sqrt{\frac{0.1097(1-0.1097)}{6000}} = (0.1018, 0.1176)$$

- 95% confidence interval for proportion
- So 95% certain processor spends 10.2-11.8% of its time in OS

51

Number of measurements for proportions

$$(1-e)\bar{p} = \bar{p} - z_{1-\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$e\bar{p} = z_{1-\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$n = \frac{z_{1-\alpha/2}^2\bar{p}(1-\bar{p})}{(e\bar{p})^2}$$

52

Number of measurements for proportions

- How long to run OS experiment?
- Want 95% confidence
- ± 0.5%

53

Number of measurements for proportions

- How long to run OS experiment?
- Want 95% confidence
- ± 0.5%
- *e = 0.005*
- *p = 0.1097*

54

Number of measurements for proportions

$$n = \frac{z_{1-\alpha/2}^2 \bar{p}(1-\bar{p})}{(e\bar{p})^2}$$

$$= \frac{(1.960)^2(0.1097)(1-0.1097)}{\left[0.005(0.1097)\right]^2}$$

$$= 1{,}247{,}102$$

- 10 ms interrupts

→ 3.46 hours

55

Confidence Interval Estimation for Variances

Confidence Interval for the Variance

- If the original variable is normally distributed then the chi-square distribution can be used to develop a confidence interval estimate of the population variance.
- The (1-α)% confidence interval for σ^2 is

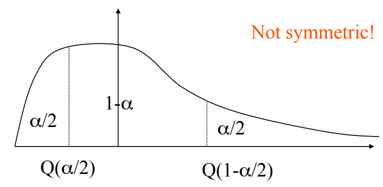
$$\frac{(n-1)s^2}{\chi^2_U} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_L}$$

χ^2_L : lower critical value of χ^2

χ^2_U : upper critical value of χ^2

57

Chi-square distribution



58

95% confidence interval for the population variance for a sample of size 100 for a N(3,2) population.

2.91903	average
4.71435	variance
2.17126	std deviation
73.36110	lower critical value of chi-square for 95%
128.42193	upper critical value of chi-square for 95%

lower bound for confidence interval for the variance	3.634277
upper bound for confidence interval for the variance	6.361966

In Excel:
CHIINV (0.975, 99) (1-α/2)
CHIINV (0.025, 99) (α/2)

The population variance (4 in this case) is in the interval (3.6343, 6.362) with 95% confidence.

59

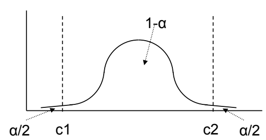
Confidence Interval for the Variance

If the population is not normally distributed, the confidence interval, especially for small samples, is not very accurate.

60

Key Assumption

- Measurement errors are Normally distributed.
- Is this true for most measurements on real computer systems?



61

Key Assumption

- Saved by the Central Limit Theorem
  Sum of a "large number" of values from any distribution will be Normally (Gaussian) distributed.
- What is a "large number?"
  - Typically assumed to be >≈ 6 or 7.

62

Normalizing data for confidence intervals

- If the underlying distribution of the data being measured is not normal, then the data must be ***normalized***
  - Find the arithmetic mean of four or more randomly selected measurements
  - Find confidence intervals for the means of these average values
    - We can no longer obtain a confidence interval for the individual values
    - Variance for the aggregated events tends to be smaller than the variance of the individual events

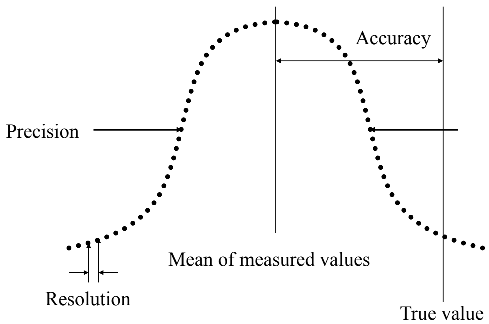
63

Summary

- Use statistics to
  - Deal with noisy measurements
  - Estimate the true value from sample data
- Errors in measurements are due to:
  - Accuracy, precision, resolution of tools
  - Other sources of noise
  - → Systematic, random errors

64

Summary (cont'd): Model errors with bell curve



65

Summary (cont'd)

- Use confidence intervals *to quantify precision*
- Confidence intervals for
  - Mean of *n* samples
  - Proportions
  - Variance
- Confidence level
  - Pr(population parameter within computed interval)
- Compute number of measurements needed for desired interval width

66