Variability in Data
CS 239
Experimental Methodologies for System Software
Peter Reiher
April 10, 2007

Introduction

- Summarizing variability in a data set
- Estimating variability in sample data

Summarizing Variability

- A single number rarely tells the entire story of a data set
- Usually, you need to know how much the rest of the data set varies from that index of central tendency

Why Is Variability Important?

- Consider two Web servers -
- Server A services all requests in 1 second
- Server B services 90% of all requests in .5 seconds
- But 10% in 55 seconds
- Both have mean service times of 1 second
- But which would you prefer to use?

Indices of Dispersion

- Measures of how much a data set varies
  - Range
  - Variance and standard deviation
  - Percentiles
  - Semi-interquartile range
  - Mean absolute deviation

Range

- Minimum and maximum values in data set
- Can be kept track of as data values arrive
- Variability characterized by difference between minimum and maximum
- Often not useful, due to outliers
- Minimum tends to go to zero
- Maximum tends to increase over time
- Not useful for unbounded variables

Example of Range

- For data set:

  2, 5.4, -17, 2056, 445, -4.8, 84.3, 92, 27, -10
- Maximum is 2056
- Minimum is -17
- Range is 2073
- While arithmetic mean is 268

Variance (and Its Cousins)

- Sample variance is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

- Variance is expressed in units of the measured quantity squared
  - Which isn't always easy to understand
- Standard deviation and the coefficient of variation are derived from variance

Variance Example

- For data set

  2, 5.4, -17, 2056, 445, -4.8, 84.3, 92, 27, -10
- Variance is 413746.6
- Given a mean of 268, what does that variance indicate?

Standard Deviation

- The square root of the variance
- In the same units as the units of the metric
- So easier to compare to the metric

Standard Deviation Example

- For data set

  2, 5.4, -17, 2056, 445, -4.8, 84.3, 92, 27, -10
- Standard deviation is 643
- Given a mean of 268, clearly the standard deviation shows a lot of variability from the mean

Coefficient of Variation

- The ratio of the mean and standard deviation
- Normalizes the units of these quantities into a ratio or percentage
- Often abbreviated C.O.V.

Coefficient of Variation Example

- For data set

  2, 5.4, -17, 2056, 445, -4.8, 84.3, 92, 27, -10
- Standard deviation is 643
- The mean of 268
- So the C.O.V. is 643/268 = 2.4

Percentiles

- Specification of how observations fall into buckets
- E.g., the 5-percentile is the observation that is at the lower 5% of the set
- The 95-percentile is the observation at the 95% boundary of the set
- Useful even for unbounded variables

Relatives of Percentiles

- Quantiles - fraction between 0 and 1
  - Instead of percentage
  - Also called fractiles
- Deciles - percentiles at the 10% boundaries
  - First is 10-percentile, second is 20-percentile, etc.
- Quartiles - divide data set into four parts
  - 25% of sample below first quartile, etc.
  - Second quartile is also the median

Calculating Quantiles

- The ? -quantile is estimated by sorting the set
- Then take the [(n-1)? +1]th element
  - Rounding to the nearest integer index

Quartile Example

- For data set

  2, 5.4, -17, 2056, 445, -4.8, 84.3, 92, 27, -10
  - (10 observations)
- Sort it:

  -17, -10, -4.8, 2, 5.4, 27, 84.3, 92, 445, 2056
- The first quartile Q_1 is -4.8
- The third quartile Q_3 is 92

Interquartile Range

- Yet another measure of dispersion
- The difference between Q3 and Q1
- Semi-interquartile range -

$$SIQR \ ? \ \frac{Q_3 \ ? \ Q_1}{2}$$

- Often interesting measure of what's going on in the middle of the range

Semi-Interquartile Range Example

- For data set

  -17, -10, -4.8, 2, 5.4, 27, 84.3, 92, 445, 2056

- Q_3 is 92
- Q_1 is -4.8

  $$SIQR = \frac{Q_3 - Q_1}{2} = \frac{92 - (-4.8)}{2} = 48$$

- So outliers cause much of variability

Mean Absolute Deviation

- Another measure of variability
- Mean absolute deviation = $\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$
- Doesn’t require multiplication or square roots

Mean Absolute Deviation Example

- For data set

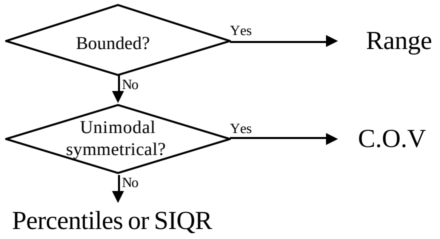
  -17, -10, -4.8, 2, 5.4, 27, 84.3, 92, 445, 2056

- Mean absolute deviation = $\frac{1}{10}\sum_{i=1}^{10}|x_i - \bar{x}|$
- Or 393

Sensitivity To Outliers

- From most to least,
  - Range
  - Variance
  - Mean absolute deviation
  - Semi-interquartile range

So, Which Index of Dispersion Should I Use?



- But always remember what you’re looking for

Determining Distributions for Datasets

- If a data set has a common distribution, that’s the best way to summarize it
- Saying a data set is uniformly distributed is more informative than just giving its mean and standard deviation

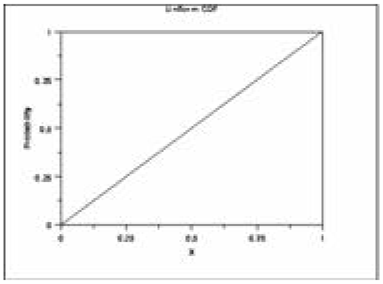
Some Commonly Used Distributions

- Uniform distribution
- Normal distribution
- Exponential distribution
- There are many others

Uniform Distribution

- All values in a given range are equally likely
- Often normalized to a range from zero to one
- Suggests randomness in phenomenon being tested
  - Pdf: $f(x)\ ?\ \frac{1}{B\ ?\ A}$
  - CDF: $f(x)\ ?\ x$
    - Assuming $0\ ?\ x\ ?\ 1$

CDF for Uniform Distribution



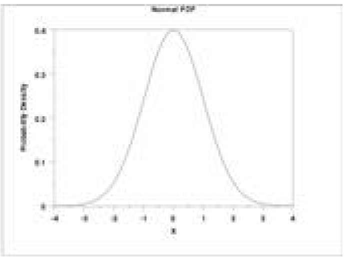
Normal Distribution

- Some value of random variable is most likely
  - Declining probabilities of values as one moves away from this value
  - Equally on either side of most probable value
- Extremely widely used
- Generally sort of a “default distribution”
  - Which isn’t always right . . .

PDF and CDF for Normal Distribution

- PDF expressed in terms of
  - Location parameter μ (the popular value)
  - Scale parameter s (how much spread)
  - PDF is $$f(x)\ ?\ \frac{e^{?(x?\ ?)^2/(2?^2)}}{?\sqrt{2?}}$$
  - CDF doesn’t exist in closed form

PDF for Normal Distribution



Exponential Distribution

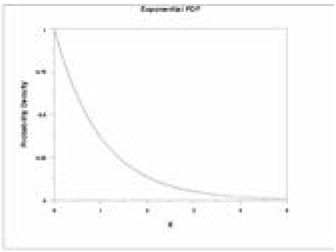
- Describes value that declines over time
  - E.g., failure probabilities
  - Described in terms of location parameter *μ*
  - And scale parameter *ß*
  - Standard exponential when *μ*= 0 and *ß*=1
- PDF:

$$f(x) ? \frac{1}{?} e^{?(x? ?)/ ?} \qquad f(x) ? e^{? x} \text{ for } \mu = 0 \text{ and } ß=1$$

- CDF:

$$f(x) ? 1 ? e^{? x/ ?}$$

PDF of Exponential Distribution



Methods of Determining a Distribution

- So how do we determine if a data set matches a distribution?
  - Plot a histogram
  - Quantile-quantile plot
  - Statistical methods (not covered in this class)

Plotting a Histogram

- Suitable if you have a relatively large number of data points

1. Determine range of observations
2. Divide range into buckets
3. Count number of observations in each bucket
4. Divide by total number of observations and plot it as column chart

Problem With Histogram Approach

- Determining cell size
  - If too small, too few observations per cell
  - If too large, no useful details in plot
- If fewer than five observations in a cell, cell size is too small

Quantile-Quantile Plots

- More suitable for small data sets
- Basically, guess a distribution
- Plot where quantiles of data theoretically should fall in that distribution
  - Against where they actually fall
- If plot is close to linear, data closely matches that distribution

Obtaining Theoretical Quantiles

- Must determine where the quantiles should fall for a particular distribution
- Requires inverting distribution's CDF
  - Then determining quantiles for observed points
  - Then plugging in quantiles to inverted CDF

Inverting a Distribution

- Many common distributions have already been inverted
  - How convenient
- For others that are hard to invert, tables and approximations are often available
  - Nearly as convenient

Is Our Sample Data Set Normally Distributed?

- Our data set was

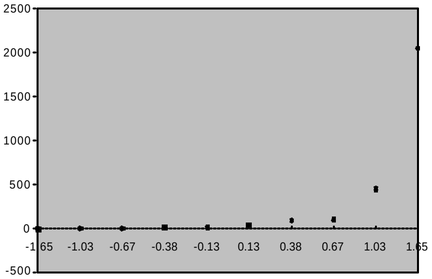
  -17, -10, -4.8, 2, 5.4, 27, 84.3, 92, 445, 2056
- Does this match the normal distribution?
- The normal distribution doesn't invert nicely
- But there is an approximation:

$$x_i \; ? \; 4.91 ? q_i^{0.14} \; ? \; ?1 \; ? \; q_i \, ?^{0.14} \, ?$$

Data For Example Normal Quantile-Quantile Plot

i	q_i	y_i	x_i
1	0.05	-17	-1.64684
2	0.15	-10	-1.03481
3	0.25	-4.8	-0.67234
4	0.35	2	-0.38375
5	0.45	5.4	-0.1251
6	0.55	27	0.1251
7	0.65	84.3	0.383753
8	0.75	92	0.672345
9	0.85	445	1.034812
10	0.95	2056	1.646839

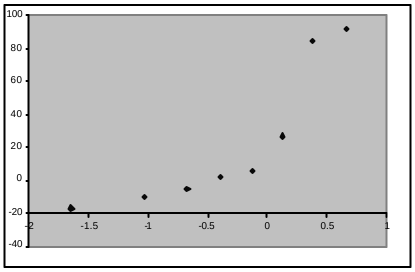
Example Normal Quantile-Quantile Plot



Analysis

- Well, it ain't normal
  - Because it isn't linear
  - Tail at high end is too long for normal
- But perhaps the lower part of the graph is normal?

Quantile-Quantile Plot of Partial Data



Partial Data Plot Analysis

- Doesn’t look particularly good at this scale, either
- OK for first five points
- Not so OK for later ones

Samples

- How tall is a human?
  - Could measure every person in the world
  - Or could measure every person in this room
- Population has *parameters*
  - Real and meaningful
- Sample has *statistics*
  - Drawn from population
  - Inherently erroneous

Sample Statistics

- How tall is a human?
  - People in Haines A82 have a mean height
  - People in BH 3564 have a different mean
- Sample mean is itself a random variable
  - Has own distribution

Estimating Population from Samples

- How tall is a human?
  - Measure everybody in this room
  - Calculate sample mean $\bar{x}$
  - Assume population mean ? equals $\bar{x}$
- But we didn’t test everyone, so that’s probably not quite right
- What is the error in our estimate?

Estimating Error

- Sample mean is a random variable
  ? Sample mean has some distribution
  ? Multiple sample means have “mean of means”
- Knowing distribution of means can estimate error

Estimating Value of a Random Variable

- How tall is Fred?
- Suppose average human height is 170 cm
  - ? Fred is 170 cm tall
  - – Yeah, right
- Safer to assume a range

Confidence Intervals

- How tall is Fred?
  - – Suppose 90% of humans are between 155 and 190 cm
  - ? Fred is between 155 and 190 cm
- We are *90% confident* that Fred is between 155 and 190 cm

Confidence Interval of Sample Mean

- Knowing where 90% of sample means fall we can state a 90% confidence interval
- Key is *Central Limit Theorem:*
  - – Sample means are normally distributed
  - – Only if independent
  - – Mean of sample means is population mean ?
  - – Standard deviation (*standard error*) is $^{?}\!/\!\sqrt{n}$

Estimating Confidence Intervals

- Two formulas for confidence intervals
  - – Over 30 samples from any distribution: *z*-distribution
  - – Small sample from normally distributed population: *t*-distribution
- Common error: using *t*-distribution for non-normal population

The *z* Distribution

- Interval on either side of mean:

$$\bar{x}\ ?\ z_{1??/2} ? \frac{s}{\sqrt{n}} ?$$

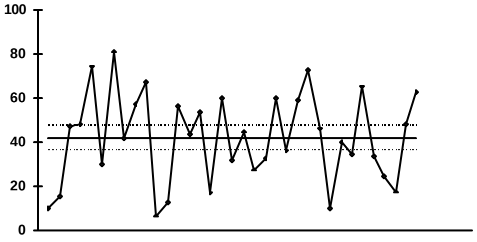
- Significance level ? is small for large confidence levels
- Tables are tricky: be careful!

Example of *z* Distribution

- 35 samples:
  10 16 47 48 74 30 81 42 57 67 7 13 56 44 54 17 60 32 45 28 33 60 36 59 73 46 10 40 35 65 34 25 18 48 63
- Sample mean $\bar{x}$ = 42.1
- Standard deviation *s* = 20.1
- n = 35
- 90% confidence interval:

$$42.1\ ?\ (1.645)\frac{20.1}{\sqrt{35}}\ ?\ (36.5, 47.7)$$

Graph of *z* Distribution Example



The *t* Distribution

- Formula is almost the same:

$$\bar{x} ? t_{1?\frac{?}{2};n?1} ? \frac{s}{\sqrt{n}} ?$$

- Usable only for normally distributed populations!
- But works with small samples

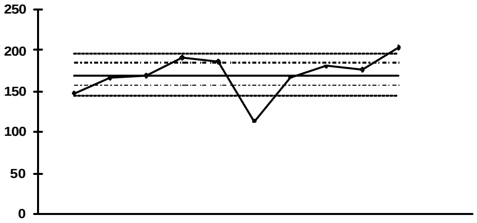
Example of *t* Distribution

- 10 height samples: 148 166 170 191 187 114 168 180 177 204
- Sample mean $\bar{x}$ = 170.5, standard deviation *s* = 25.1, n = 10
- 90% confidence interval is

$$170.5 ? (1.833)\frac{25.1}{\sqrt{10}} ? (156.0, 185.0)$$

- 99% interval is (144.7, 196.3)

Graph of *t* Distribution Example



Getting More Confidence

- Asking for a higher confidence level widens the confidence interval
- How tall is Fred?
  - 90% sure he's between 155 and 190 cm
  - We want to be 99% sure we're right
  - So we need more room: 99% sure he's between 145 and 200 cm

Making Decisions

- Why do we use confidence intervals?
  - Summarizes error in sample mean
  - Gives way to decide if measurement is meaningful
  - Allows comparisons in face of error
- But remember: at 90% confidence, 10% of sample means *do not* include population mean
- And confidence intervals apply to means, not individual data readings

Testing for Zero Mean

- Is population mean significantly nonzero?
- If confidence interval includes 0, answer is *no*
- Can test for any value (mean of sums is sum of means)
- Example: our height samples are consistent with average height of 170 cm
  - Also consistent with 160 and 180!

Comparing Alternatives

- Often need to find better system
  - Choose fastest computer to buy
  - Prove our algorithm runs faster
- Different methods for paired/unpaired observations
  - *Paired* if *i*th test on each system was same
  - *Unpaired* otherwise

Comparing Paired Observations

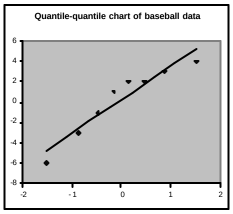
- For each test calculate performance difference
- Calculate confidence interval for mean of differences
- If interval includes zero, systems aren't different
  - If not, sign indicates which is better

Example: Comparing Paired Observations

- Do home baseball teams outscore visitors?
- Sample from 4-7-07:
  - H 1 8 5 5 5 7 3 1
  - V 7 5 3 6 1 5 2 4
  - H-V -6 3 2 -1 4 2 1 -3
- Assume a normal population for the moment
  - n = 8, Mean = .25, s= 3.37, 90% interval (-2, 2.5)
  - Can't tell from this data

Was the Data Normally Distributed?

- Check by plotting quantile-quantile chart
- Pretty good fit to the line
- So the normal assumption is plausible



Comparing Unpaired Observations

- Start with confidence intervals for each sample
  - If no overlap:
    - Systems are different and higher mean is better (for HB metrics)
  - If overlap and each CI contains other mean:
    - Systems are not different at this level
    - If close call, could lower confidence level
  - If overlap and one mean isn't in other CI
    - Must do ***t**-test*

The *t*-test (1)

1. Compute sample means $\overline{x}_a$ and $\overline{x}_b$
2. Compute sample standard deviations s_a and s_b
3. Compute mean difference = $\overline{x}_a ? \overline{x}_b$
4. Compute standard deviation of difference: $s ? \sqrt{\frac{s_a^2}{n_a} ? \frac{s_b^2}{n_b}}$

The *t*-test (2)

5. Compute effective degrees of freedom:

$$? ? \frac{?s_a^2 / n_a ? s_b^2 / n_b?^2}{\frac{1}{n_a ? 1}?\frac{s_a^2}{n_a}?^2 ? \frac{1}{n_b ? 1}?\frac{s_b^2}{n_b}?^2} ? 2$$

6. Compute the confidence interval:

$$?\overline{x}_a ? \overline{x}_b? □ t_{?1? ?/2;? ?}s$$

7. If interval includes zero, no difference

Comparing Proportions

- If *k* of *n* trials give a certain result, then confidence interval is

$$\frac{k}{n} □ z_{1??/2}\frac{\sqrt{k ? k^2 / n}}{n}$$

- If interval includes 0.5, can't say which outcome is statistically meaningful
- Must have k>10 to get valid results

Special Considerations

- Selecting a confidence level
- Hypothesis testing
- One-sided confidence intervals
- Estimating required sample size

Selecting a Confidence Level

- Depends on cost of being wrong
- 90%, 95% are common values for scientific papers
- Generally, use highest value that lets you make a firm statement
  - But it's better to be consistent throughout a given paper

Hypothesis Testing

- The *null hypothesis* (H_0) is common in statistics
  - Confusing due to double negative
  - Gives less information than confidence interval
  - Often harder to compute
- Should understand that rejecting null hypothesis implies result is meaningful

One-Sided Confidence Intervals

- Two-sided intervals test for mean being outside a certain range (see “error bands” in previous graphs)
- One-sided tests useful if only interested in one limit
- Use $z_{1-?}$ or $t_{1-?;n}$ instead of $z_{1-?/2}$ or $t_{1-?/2;n}$ in formulas

Sample Sizes

- Bigger sample sizes give narrower intervals
  - Smaller values of t, v as n increases
  - $\sqrt{n}$ in formulas
- But sample collection is often expensive
  - What is the minimum we can get away with?

How To Estimate Sample Size

- Take a small number of measurements
- Use statistical properties of the small set to estimate required size
- Based on desired confidence of being within some percent of true mean
- Gives you a confidence interval of a certain size
  - At a certain confidence that you’re right

Choosing a Sample Size

- To get a given percentage error ±r%:

$$n\ ?\ ?\frac{100zs}{r\bar{x}}?^{2}$$

- Here, z represents either z or t as appropriate

Example of Choosing Sample Size

- Five runs of a compilation took 22.5, 19.8, 21.1, 26.7, 20.2 seconds
- How many runs to get ±5% confidence interval at 90% confidence level?
- $\bar{x}$ = 22.1, s = 2.8, $t_{0.95;4}$ = 2.132

$$n\ ?\ ?\frac{??100??2.132??2.8??}{?5??22.1?}?^{2}\ ?\ 5.4^{2}\ ?\ 29.2$$

What Does This really Mean?

- After running five tests
- If I run a total of 30 tests
- My confidence intervals will be within 5% of the mean
- At a 90% cnfidence level