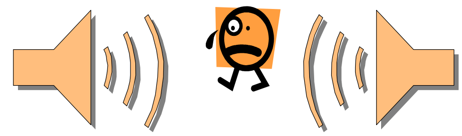
CS533
Modeling and Performance Evaluation of Network and Computer Systems

Statistics for Performance Evaluation

(Chapters 12-15)

Why do we need statistics?

1. Noise, noise, noise, noise, noise!



OK – not really this type of noise

Why Do We Need Statistics?

2. Aggregate data into meaningful information.

445 446 397 226
388 3445 188 1002
47762 432 54 12
98 345 2245 8839
77492 472 565 999
1 34 882 545 4022
827 572 597 364

$\overline{x} = \ldots$

Why Do We Need Statistics?

"Impossible things usually don't happen."
 - *Sam Treiman, Princeton University*

- *Statistics* helps us quantify "usually."

What is a *Statistic*?

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

→ A single number used to summarize a larger collection of values.

What are *Statistics*?

- "Lies, damn lies, and statistics!"
- "A collection of quantitative data."
- "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.

Objectives

- Provide intuitive conceptual background for some standard statistical tools.
  - Draw meaningful conclusions in presence of noisy measurements.
  - Allow you to correctly and intelligently apply techniques in new situations.

→ Don't simply plug and crank from a formula!

Outline

- Introduction
- Basics
- Indices of Central Tendency
- Indices of Dispersion
- Comparing Systems
- Misc
- Regression
- ANOVA

Basics (1 of 3)

- Independent Events:
  - One event does not affect the other
  - Knowing probability of one event does not change estimate of another
- Cumulative Distribution (or Density) Function:
  - $F_x(a) = P(x<=a)$
- Mean (or Expected Value):
  - Mean $\mu = E(x) = \Sigma(p_i x_i)$ for i over n
- Variance:
  - Square of the distance between x and the mean
    - $(x- \mu)^2$
  - $Var(x) = E[(x- \mu)^2] = \Sigma p_i (x_i - \mu)^2$
  - Variance is often σ. Square root of variance, σ^2, is standard deviation

Basics (2 of 3)

- Coefficient of Variation:
  - Ratio of standard deviation to mean
  - C.O.V. = σ / μ
- Covariance:
  - Degree two random variables vary with each other
  - Cov = $\sigma^2_{xy} = E[(x- \mu_x)(y- \mu_y)]$
  - Two independent variables have Cov of 0
- Correlation:
  - Normalized Cov (between –1 and 1)
  - $\rho_{xy} = \sigma^2_{xy} / \sigma_x \sigma_y$
  - Represents degree of linear relationship

Basics (3 of 3)

- Quantile:
  - The x value of the CDF at α
  - Denoted x_α, so $F(x_\alpha) = \alpha$
  - Often want .25, .50, .75
- Median:
  - The 50-percentile (or, .5-quantile)
- Mode:
  - The most likely value of x_i
- Normal Distribution
  - Most common distribution used, "bell" curve

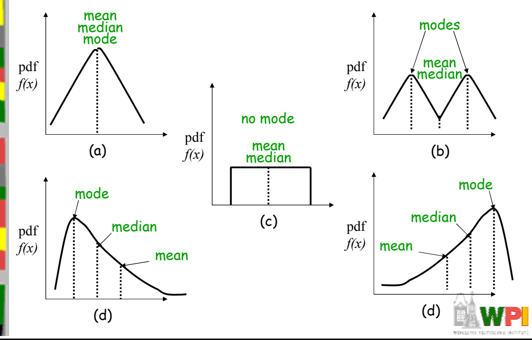
Outline

- Introduction
- Basics
- Indices of Central Tendency
- Indices of Dispersion
- Comparing Systems
- Misc
- Regression
- ANOVA

Summarizing Data by a Single Number

- *Indices of central tendency*
- Three popular: mean, median, mode
- Mean – sum all observations, divide by num
- Median – sort in increasing order, take middle
- Mode – plot histogram and take largest bucket
- Mean can be affected by outliers, while median or mode ignore lots of info
- Mean has additive properties (mean of a sum is the sum of the means), but not median or mode

Relationship Between Mean, Median, Mode



Guidelines in Selecting Index of Central Tendency

- Is it categorical?
  - → yes, use mode
    - Ex: most frequent microprocessor
- Is total of interest?
  - → yes, use mean
    - Ex: total CPU time for query (yes)
    - Ex: number of windows on screen in query (no)
- Is distribution skewed?
  - → yes, use median
  - → no, use mean

Examples for Index of Central Tendency Selection

- Most used resource in a system?
  - Categorical, so use mode
- Response time?
  - Total is of interest, so use mean
- Load on a computer?
  - Probably highly skewed, so use median
- Average configuration of number of disks, amount of memory, speed of network?
  - Probably skewed, so use median

Common Misuses of Means (1 of 2)

- Using mean of significantly different values
  - Just because mean is right, does not say it is useful
    - Ex: two samples of response time, 10 ms and 1000 ms. Mean is 505 ms but useless.
- Using mean without regard to skew
  - Does not well-represent data if skewed
    - Ex: sys A: 10, 9, 11, 10, 10 (mean 10, mode 10)
    - Ex: sys B: 5, 5, 5, 4, 31 (mean 10, mode 5)

Common Misuses of Means (2 of 2)

- Multiplying means
  - Mean of product equals product of means if two variables are independent. But:
    - if x,y are correlated E(xy) != E(x)E(y)
  - Ex: mean users system 23, mean processes per user is 2. What is the mean system processes? Not 46!

  → Processes determined by load, so when load high then users have fewer. Instead, must measure total processes and average.
- Mean of ratio with different bases (later)

Geometric Mean (1 of 2)

- Previous mean was *arithmetic mean*
  - Used when sum of samples is of interest
  - *Geometric mean* when product is of interest
- Multiply n values $\{x_1, x_2, \ldots, x_n\}$ and take n^{th} root:

  $x = (\prod x_i)^{1/n}$
- Example: measure time of network layer improvement, where 2x layer 1 and 2x layer 2 equals 4x improvement.
- Layer 7 improves 18%, 6 13%, 5, 11%, 4 8%, 3 10%, 2 28%, 1 5%
- So, geometric mean per layer:
  - $[(1.18)(1.13)(1.11)(1.08)(1.10)(1.28)(1.05)]^{1/7} - 1$
  - Average improvement per layer is 0.13, or 13%

Geometric Mean (2 of 2)

- Other examples of metrics that work in a multiplicative manner:
  - Cache hit ratios over several levels
    - And cache miss ratios
  - Percentage of performance improvement between successive versions
  - Average error rate per hop on a multi-hop path in a network

Harmonic Mean (1 of 2)

- Harmonic mean of samples $\{x_1, x_2, \ldots, x_n\}$ is:

  $n / (1/x_1 + 1/x_2 + \ldots + 1/x_n)$
- Use when arithmetic mean works for 1/x
- Ex: measurement of elapsed processor benchmark of m instructions. The ith takes t_i seconds. MIPS x_i is m/t_i
  - Since sum of instructions matters, can use harmonic mean

  $= n / [1/(m/t_1) + 1/(m/t_2) + \ldots + 1/(m/t_n)]$

  $= m / [(1/n)(t_1 + t_2 + \ldots + t_n)$

Harmonic Mean (2 of 2)

- Ex: if different benchmarks (m_i), then sum of m_i/t_i does not make sense
- Instead, use weighted harmonic mean

  $n / (w_1/x_1 + w_2/x_2 + \ldots + w_3/x_n)$
  - where $w_1 + w_2 + .. + w_n = 1$
- In example, perhaps choose weights proportional to size of benchmarks
  - $w_i = m_i / (m_1 + m_2 + .. + m_n)$
- So, weighted harmonic mean

  $(m_1 + m_2 + .. + m_n) / (t_1 + t_2 + .. + t_n)$
  - Reasonable, since top is total size and bottom is total time

Mean of a Ratio (1 of 2)

- Set of n ratios, how to summarize?
- Here, if sum of numerators and sum of denominators both have meaning, the average ratio is the ratio of averages

  Average$(a_1/b_1, a_2/b_2, \ldots, a_n/b_n)$

  $= (a_1 + a_2 + \ldots + a_n) / (b_1 + b_2 + \ldots + b_n)$

  $= [(\Sigma a_i)/n] / [(\Sigma b_i)/n]$
- Commonly used in computing mean resource utilization (example next)

Mean of a Ratio (2 of 2)

- CPU utilization:
  - For duration 1 busy 45%, 1 %45, 1 45%, 1 45%, 100 20%
  - Sum 200%, mean != 200/5 or 40%
    - The base denominators (duration) are not comparable
  - mean = sum of CPU busy / sum of durations

  = (.45+.45+.45+.45+20) / (1+1+1+1+100)

  = 21%

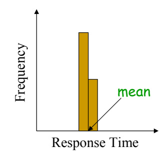
Outline

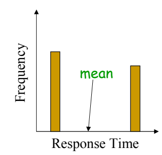
- Introduction
- Basics
- Indices of Central Tendency
- Indices of Dispersion
- Comparing Systems
- Misc
- Regression
- ANOVA

Summarizing Variability (1 of 2)

> *"Then there is the man who drowned crossing a stream with an average depth of six inches."* – W.I.E. Gates

- Summarizing by a single number is rarely enough → need statement about *variability*
  - If two systems have same mean, tend to prefer one with less variability





Summarizing Variability (2 of 2)

- *Indices of Dispersion*
  - *Range* – min and max values observed
  - *Variance* or *standard deviation*
  - 10- and 90-*percentiles*
  - (*Semi-)interquartile* range
  - *Mean absolute deviation*

(Talk about each next)

Range

- Easy to keep track of
- Record max and min, subtract
- Mostly, not very useful:
  - Minimum may be zero
  - Maximum can be from outlier
    - System event not related to phenomena studied
  - Maximum gets larger with more samples, so no "stable" point
- However, if system is bounded, for large sample, range may give bounds

Sample Variance

- Sample *variance* (can drop word "sample" if meansing is clear)
  - $s^2 = [1/(n-1)] \Sigma(x_i - \underline{x})^2$
- Notice (n-1) since only n-1 are independent
  - Also called *degrees of freedom*
- Main problem is in units squared so changing the units changes the answer squared
  - Ex: response times of .5, .4, .6 seconds

  Variance = 0.01 seconds squared or 10000 msecs squared

Standard Deviation

- So, use *standard deviation*
  - s = sqrt(s^2)
  - Same unit as *mean*, so can compare to *mean*
- Ex: response times of .5, .4, .6 seconds
  - stddev .1 seconds or 100 msecs
  - Can compare each to mean
- Ratio of *standard deviation* to *mean*?
  - Called the *Coefficient of Variation* (C.O.V.)
  - Takes units out and shows magnitude
  - Ex: above is 1/5th (or .2) for either unit

Percentiles/Quantile

- Similar to range
- Value at express percent (or fraction)
  - 90-percentile, 0.9-quantile
  - For α-quantile, sort and take $[(n-1)\alpha+1]^{th}$
    - [] means round to nearest integer
- 25%, 50%, 75% → *quartiles* (Q1, Q2, Q3)
  - Note, Q2 is also the median
- Range of Q3 – Q1 is *interquartile* range
  - ½ of (Q3 – Q1) is *semi-interquartile* range

Mean Absolute Deviation

- $(1/n)\,\Sigma|x_i - \underline{x}|$
- Similar to standard deviation, but requires no multiplication or square root
- Does not magnify outliers as much
  - (Outliers are not squared)
- So, how susceptible are indices of dispersion to outliers?

Indices of Dispersion Summary

- Ranking of affect by outliers
  - Range — susceptible
  - Variance (standard deviation)
  - Mean absolute deviation
  - Semi-interquartile range — resistant
- Use semi-interquantile (SIQR) for index of dispersion whenever using median as index of central tendency
- Note, all only applied to quantitative data
  - For qualitative (categorical) give number of categories for a given percentile of samples

Indices of Dispersion Example

(Sorted)
CPU Time

1.9	3.9
2.7	3.9
2.8	4.1
2.8	4.1
2.8	4.2
2.9	4.2
3.1	4.4
3.1	4.5
3.2	4.5
3.2	4.8
3.3	4.9
3.4	5.1
3.6	5.1
3.7	5.3
3.8	5.6
3.9	5.9

- First, sort
- Median = $[1 + 31*.5] = 16^{th} = 3.2$
- Q1 = $1 + .31 * .25 = 9^{th} = 3.9$
- Q3 = $1 + .31*.75 = 24^{th} = 4.5$
- *SIQR* = (Q3–Q1)/2 = .65
- *Variance* = 0.898
- *Stddev* = 0.948
- *Range* = 5.9 – 1.9 = 4

Selecting Index of Dispersion

- Is distribution bounded
  - Yes? → use range
- No? Is distribution unimodal symmetric?
  - Yes? → Use C.O.V.
- No?
  - Use percentiles or SIQR
- Not hard-and-fast rules, but rather *guidelines*
  - Ex: dispersion of network load. May use range or even C.O.V. But want to accommodate 90% or 95% of load, so use percentile. Power supplies similar.

Determining Distribution of Data

- Additional summary information could be the *distribution* of the data
  - Ex: Disk I/O mean 13, variance 48. Ok. Perhaps more useful to say data is *uniformly distributed* between 1 and 25.
  - Plus, distribution useful for later simulation or analytic modeling
- How do determine distribution?
  - First, plot histogram

Histograms

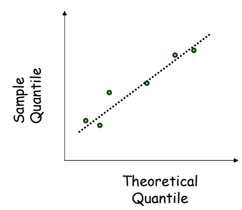
- Need: max, min, size of buckets
- Determining cell size is a problem
  - Too few, hard to see distro
  - Too many, distro lost
  - Guideline:
    - if any cell > 5 then split

Cell	#	Histogram (size 1)
1	1	X
2	5	XXXXX
3	12	XXXXXXXXXXXX
4	9	XXXXXXXXX
5	5	XXXXX

Cell	#	Histogram (size .2)
1.8	1	X
2.6	1	X
2.8	4	XXXX
3.0	2	XX
3.2	3	XXX
3.4	1	X
3.6	2	XX
3.8	4	XXXX
4.0	2	XX
4.2	2	XX
4.4	3	XXX
4.8	2	XX
5.0	2	XX
5.2	1	X
5.6	1	X
5.8	1	X

Distribution of Data

- Instead, plot observed quantile versus theoretical quantile
  - y_i is observed, x_i is theoretical
  - If distribution fits, will have line



Need to invert CDF: $q_i = F(x_i)$, or $x_i = F^{-1}(q_i)$

Where F^{-1}? Table 28.1 for many distributions

Normal distribution: $x_i = 4.91[q_i^{0.14} - (1-q_i)^{0.14}]$

Table 28.1

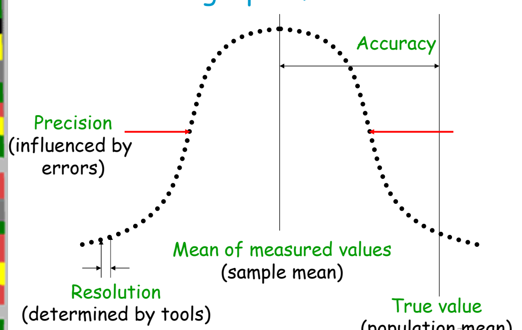
Distribution	CDF $F(x)$	Inverse
Exponential	$1 - e^{-x/a}$	$-a \ln(u)$
Extreme value	$1 - e^{-e^{\frac{x-a}{b}}}$	$a + b \ln \ln u$
Geometric	$1 - (1-p)^x$	$\left\lceil \frac{\ln(u)}{\ln(1-p)} \right\rceil$
Logistic	$1 - \frac{1}{1+e^{\frac{x-\mu}{b}}}$	$\mu - b \ln(\frac{1}{u} - 1)$
Pareto	$1 - x^{-a}$	$1/u^{1/a}$
Weibull	$1 - e^{(x/a)^b}$	$a(\ln u)^{1/b}$

Normal distribution:
$x_i = 4.91[q_i^{0.14} - (1-q_i)^{0.14}]$

Outline

- Introduction
- Basics
- Indices of Central Tendency
- Indices of Dispersion
- Comparing Systems
- Misc
- Regression
- ANOVA

Measuring Specific Values



Comparing Systems Using Sample Data

"Statistics are like alienists – they will testify for either side." – Fiorello La Guardia

- The word "sample" comes from the same root word as "example"
- Similarly, one sample does not prove a theory, but rather is an example
- Basically, a definite statement cannot be made about characteristics of all systems
- Instead, make probabilistic statement about range of most systems
  - *Confidence intervals*

Sample versus Population

- Say we generate 1-million random numbers
  - mean μ and stddev σ.
  - μ is *population mean*
- Put them in an urn draw sample of n
  - Sample $\{x_1, x_2, \ldots, x_n\}$ has mean $\underline{x}$, stddev s
- $\underline{x}$ is likely different than μ!
  - With many samples, $\underline{x}_1 \neq \underline{x}_2 \neq \ldots$
- Typically, μ is not known and may be impossible to know
  - Instead, get estimate of μ from $\underline{x}_1, \underline{x}_2, \ldots$

Confidence Interval for the Mean

- Obtain probability of μ in interval $[c_1, c_2]$
  - $\text{Prob}\{c_1 \leq \mu \leq c_2\} = 1-\alpha$
    - (c1, c2) is *confidence interval*
    - α is *significance level*
    - $100(1-\alpha)$ is *confidence level*
- Typically want α small so confidence level 90%, 95% or 99% (more later)
- Say, α =0.1. Could take k samples, find sample means, sort
  - Interval: $[1+0.05(k-1)]^{th}$ and $[1+0.95(k-1)]^{th}$
    - 90% confidence interval
- We have to take k samples, each of size n?

Central Limit Theorem

Sum of a "large" number of values from any distribution will be normally distributed.

- Do not need many samples. One will do.

  $\underline{x} \sim N(\mu, \sigma/\text{sqrt}(n))$
- Standard error = σ /sqrt(n)
  - As sample size n increases, error decreases
- So, a $100(1-\alpha)$% confidence interval for a population mean is:

  $(\underline{x}-z_{1-\alpha/2}s/\text{sqrt}(n), \underline{x}+z_{1-\alpha/2}s/\text{sqrt}(n))$
- Where $z_{1-\alpha/2}$ is a $(1-\alpha/2)$-quantile of a unit normal (Table A.2 in appendix, A.3 common)

Confidence Interval Example

(Sorted) CPU Time

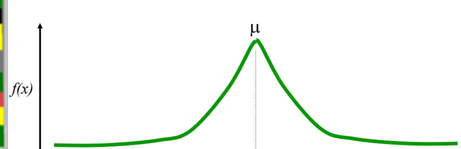
1.9	3.9
2.7	3.9
2.8	4.1
2.8	4.1
2.8	4.2
2.9	4.2
3.1	4.4
3.1	4.5
3.2	4.5
3.2	4.8
3.3	4.9
3.4	5.1
3.6	5.1
3.7	5.3
3.8	5.6
3.9	5.9

- $\underline{x}$ = 3.90, stddev s=0.95, n=32
- A 90% confidence interval for the population mean (μ):

  3.90 +- (1.645)(0.95)/sqrt(32)

  = (3.62, 4.17)
- With 90% confidence, μ in that interval. Chance of error 10%.
  - If we took 100 samples and made confidence intervals as above, in 90 cases the interval includes μ and in 10 cases would not include μ

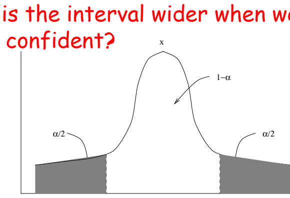
Meaning of Confidence Interval



Sample	Includes μ?
1	yes
2	yes
3	no
…	
100	yes
Total	yes $\geq 100(1-\alpha)$
Total	no $\leq 100\alpha$

How does the Interval Change?

- 90% CI = [6.5, 9.4]
  - 90% chance real value is between 6.5, 9.4
- 95% CI = [6.1, 9.7]
  - 95% chance real value is between 6.1, 9.7
- Why is the interval wider when we are more confident?



What if n not large?

- Above only applies for large samples, 30+
- For smaller n, can only construct confidence intervals if observations come from normally distributed population
  - Is that true for computer systems?

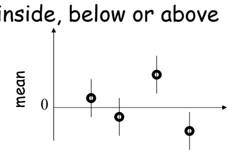
$(\underline{x}-t_{[1-\alpha/2;n-1]}s/\text{sqrt}(n),\ \underline{x}+t_{[1-\alpha/2;n-1]}s/\text{sqrt}(n))$

- Table A.4. (Student's t distribution. "Student" was an anonymous name)

Again, n-1 degrees freedom

Testing for a Zero Mean

- Common to check if a measured value is significantly different than zero
- Can use confidence interval and then check if 0 is inside interval.
- May be inside, below or above



Note, can extend this to include testing for different than any value a

Example: Testing for a Zero Mean

- Seven workloads
- Difference in CPU times of two algorithms
  {1.5, 2.6, -1.8, 1.3,-0.5, 1.7, 2.4}
- Can we say with 99% confidence that one algorithm is superior to another?
- $n = 7$, $\alpha = 0.01$
- mean = 7.20/7 = 1.03
- variance = 2.57 so stddev = sqrt(2.57) = 1.60
- CI = 1.03 +- tx1.60/sqrt(7) = 1.03 +- 0.605t
- 1 - α/2 = .995, so t[0.995;6] = 3.707 (Table A.4)
- 99% confidence interval = (-1.21, 3.27)

→ With 99% confidence, algorithm performances are identical

Comparing Two Alternatives

- Often want to compare system
  - System A with system B
  - System "before" and system "after"
- Paired Observations
- Unpaired Observations
- Approximate Visual Test

Paired Observations

- If n experiments such that 1-to-1 correspondence from test on A with test on B then *paired*
  - (If no correspondence, then *unpaired*)
- Treat two samples as one sample of n pairs
- For each pair, compute difference
- Construct confidence interval for difference
- If CI includes zero, then systems are not significantly different

Example: Paired Observations

- Measure different size workloads on A and B
  {(5.4, 19.1), (16.6, 3.5), (0.6,3.4), (1.4,2.5), (0.6, 3.6) (7.3, 1.7)}
- Is one system better than another?
- Six observed differences
  - {-13.7, 13.1, -2.8, -1.1, -3.0, 5.6}
- Mean = -.32, stddev = 9.03
- CI = -0.32 +- t[sqrt(81.62/6)] = -0.32 +- t(3.69)
- The .95 quantile of t with 5 degrees of freedom = 2.015
- 90% confidence interval = (-7.75, 7.11)
- Therefore, two systems not different

Unpaired Observations

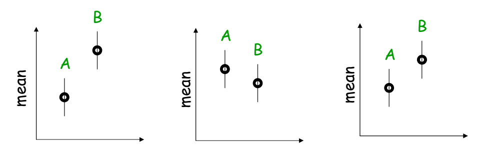
- Systems A, B with samples n_a and n_b
- Compute sample means: $\underline{x}_a$, $\underline{x}_b$
- Compute standard devs: s_a, s_b
- Compute mean difference: $\underline{x}_a$-$\underline{x}_b$
- Compute stddev of mean difference:
  - S = sqrt($s_a^2/n_a + s_b^2/n_b$)
- Compute effective degrees of freedom
- Compute confidence interval
- If interval includes zero, not a significant difference

Example: Unpaired Observations

- Processor time for task on two systems
  - A: {5.36, 16.57, 0.62, 1.41, 0.64, 7.26}
  - B: {19.12, 3.52, 3.38, 2.50, 3.60, 1.74}
- Are the two systems significantly different?
- Mean x_a = 5.31, s_a^2 = 37.92, n_a=6
- Mean x_b = 5.64, s_b^2 = 44.11, n_b =6
- Mean difference x_a-x_b = -0.33
- Stddev of mean difference = 3.698
- t is 1.71
- 90% confidence interval = (-6.92, 6.26)
  - Not different

Approximate Visual Test

- Compute confidence interval for means
- See if they overlap



CIs do not overlap → A higher than B

CIs do overlap and Mean of one in another → Not different

CIs do overlap but mean of one not in another → Do t test

Example: Approximate Visual Test

- Processor time for task on two systems
  - A: {5.36, 16.57, 0.62, 1.41, 0.64, 7.26}
  - B: {19.12, 3.52, 3.38, 2.50, 3.60, 1.74}
- t-value at 90%, 5 is 2.015
- 90% confidence intervals
  - A = 5.31 +-(2.015)sqrt(37.92/6) = (0.24,10.38)
  - B = 5.64 +-(2.015)sqrt(44.11/6) = (0.18,11.10)
- The two confidence intervals overlap and the mean of one falls in the interval of another. Therefore the two systems are not different without unpaired t test

Outline

- Introduction
- Basics
- Indices of Central Tendency
- Indices of Dispersion
- Comparing Systems
- Misc
- Regression
- ANOVA

What Confidence Level to Use?

- Often see 90% or 95% (or even 99%)
- Choice is based on loss if population parameter is outside or gain if parameter inside
  - If loss is high compared to gain, use high confidence
  - If loss is low compared to gain, use low confidence
  - If loss is negligible, low is fine
- Example:
  - Lottery ticket $1, pays $5 million
  - Chance of winning is 10^{-7} (1 in 10 million)
  - To win with 90% confidence, need 9 million tickets
    - No one would buy that many tickets!
  - So, most people happy with 0.01% confidence

Hypothesis Testing

- Most stats books have a whole chapter
- Hypothesis test usually accepts/rejects
  - Can do that with confidence intervals
- Plus, interval tells us more … precision
- Ex: systems A and B
  - CI (-100,100) we can say "no difference"
  - CI(-1, 1) say "no difference" loudly
- Confidence intervals easier to explain since units are the same as those being measured
  - Ex: more useful to know range 100 to 200 than that the probability of it being less than 110 is 3%

One-Sided Confidence Intervals

- At 90% confidence, 5% chance lower than limit and 5% chance higher than limit
- Sometimes, only want one-sided comparison
  - Say, test if mean is greater than value

    $(\bar{x}-t_{[1-\alpha;n-1]}s/\text{sqrt}(n), \bar{x})$
  - Use $1-\alpha$ instead of $1-\alpha/2$
- Similarly (but with +) for upper confidence limit
- Can use z-values if more than 30

Confidence Intervals for Proportions

- Categorical variables often has probability with each category → called *proportions*
  - Want CI on proportions
- Each sample of n observations gives a sample proportion (say, of type 1)
  - n_1 of n observations are type 1

    $p = n_1 / n$
- CI for p: $p+-z_{1-\alpha/2}\text{sqrt}(p(1-p)/n)$
- Only valid if $np \geq 10$
  - Otherwise, too complicated. See stats book.

Example: CI for Proportions

- 10 of 1000 pages printed are illegible

  $p = 10/1000 = 0.01$
- Since $np \geq 10$ can use previous equation

CI = p +- z(sqrt(p(1-p)/n))

= 0.01 +- z(sqrt(0.01(0.99)/1000)

= 0.01 +- 0.003z

90% CI = 0.01 +- (0.003)(1.645) = (0.005, 0.015)

- Thus, at 90% confidence we can say 0.5% to 1.5% of the pages are illegible.
  - There is a 10% chance this statement is in error

Determining Sample Size

- The larger the sample size, the higher the confidence in the conclusion
  - Tighter CIs since divided by sqrt(n)
  - But more samples takes more resources (time)
- Goal is to find the smallest sample size to provide the desired confidence in the results
- Method:
  - small set of preliminary measurements
  - use to estimate variance
  - use to determine sample size for accuracy

Sample Size for Mean

- Suppose we want mean performance with accuracy of +-r% at $100(1-\alpha)$% confidence
- Know for sample size n, CI is

  $\bar{x}$ +- z(s/sqrt(n))
- CI should be [$\bar{x}$(1-r/100), $\bar{x}$(1+r/100)]

  $\bar{x}$ +- z(s/sqrt(n)) = $\bar{x}$(1 +- r/100)

  z(s/sqrt(n)) = $\bar{x}$(r/100)

  $n = [(100zs)/(r\bar{x})]^2$

Example: Sample Size for Mean

- Preliminary test:
  - response time 20 seconds
  - stddev = 5 seconds
- How many repetitions to get response time accurate within 1 second at 95% confidence

  $\underline{x}$=20, s=5, z=1.960, r=5 (1 sec is 5% of 20)

  $n = [(100 \times 1.960 \times 5) / (5 \times 20)]^2$

  $= (9.8)^2$

  $= 96.04$
- So, a total of 97 observations are needed
- Can extend to proportions (not shown)

Example: Sample Size for Comparing Alternatives

- Need non-overlapping confidence intervals
- Algorithm A loses 0.5% of packets and B loses 0.6%
- How many packets do we need to state that alg A is better than alg B at 95%?

CI for A: $0.005 \pm 1.960[0.005(1-0.005)/n]^{\frac{1}{2}}$

CI for B: $0.006 \pm 1.960[0.006(1-0.006)/n]^{\frac{1}{2}}$

- Need upper edge of A not to overlap lower edge of B

  $0.005 + 1.960[0.005(1-0.005)/n]^{\frac{1}{2}} <$

  $0.006 - 1.960[0.006(1-0.006)/n]^{\frac{1}{2}}$

  solve for n: $n > 84{,}340$
- So, need 85000 packets

Summary

- Statistics are tools
  - Help draw conclusions
  - Summarize in a meaningful way in presence of noise
- Indices of central tendency and Indices of central dispersion
  - Summarize data with a few numbers
- Confidence intervals

Outline

- Introduction
- Basics
- Indices of Central Tendency
- Indices of Dispersion
- Comparing Systems
- Misc
- Regression
- ANOVA

Regression

> *"I see your point … and raise you a line."*
> – Elliot Smorodinksy

- Expensive (and sometimes impossible) to measure performance across all possible input values
- Instead, measure performance for limited inputs and use to produce model over range of input values
  - Build *regression* model

Linear Regression (1 of 2)

- Captures linear relationship between input values and response
  - *Least-squares minimization*
- Of the form:

  $y = a + bx$
- Where x input, y response and we want to know a and b
- If y_i is measured for input x_i, then each pair (x_i, y_i) can be written:

  $y_i = a + bx_i + e_i$
- where e_i is residual (error) for regression model

Linear Regression (2 of 2)

- The sum of the errors squared:

  $SSE = \Sigma e_i^2 = \Sigma(y_i - a - bx_i)^2$
- Find a and b that minimizes SSE
- Take derivative with respect to a and then b and then set both to zero

  $na + b\Sigma x_i = \Sigma y_i$ (1) (two equations in two unknowns)

  $a\Sigma x_i + b\Sigma x_i^2 = \Sigma x_i y_i$
- Solving for b gives:

  $$b = \frac{n\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{n\Sigma x_i^2 - (\Sigma x_i)^2}$$
- Using (1) and solving for a:

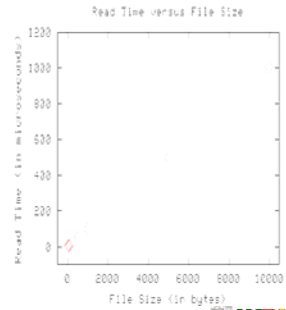
  $a = \underline{y} - b\underline{x}$

WPI

Linear Regression Example (1 of 3)

File Size (bytes)	Time (μsec)
10	3.8
50	8.1
100	11.9
500	55.6
1000	99.6
5000	500.2
10000	1006.1

Develop linear regression model for time to read file of size bytes



WPI

Linear Regression Example (2 of 3)

File Size (bytes)	Time (μsec)
10	3.8
50	8.1
100	11.9
500	55.6
1000	99.6
5000	500.2
10000	1006.1

Develop linear regression model for time to read file of size bytes

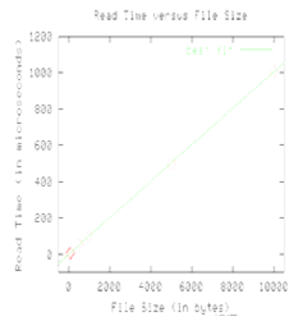
- $\Sigma x_i = 16{,}660.0$
- $\Sigma y_i = 1685.3$
- $\Sigma x_i y_i = 12{,}691{,}033.0$
- $\Sigma x_i^2 = 126{,}262{,}600.0$
- $\underline{x} = 2380$
- $\underline{y} = 240.76$
- $$b = \frac{(7)(12691033) - (16660)(1685.3)}{(7)(126262600) - (16660)^2}$$
- $a = 240.76 - .1002(2380)$

  $= 2.24$
- $y = 2.24 + 0.1002x$

WPI

Linear Regression Example (3 of 3)

File Size (bytes)	Time (μsec)
10	3.8
50	8.1
100	11.9
500	55.6
1000	99.6
5000	500.2
10000	1006.1



$y = 2.24 + 0.1002x$

Ex: predict time to read 3k file is 303 μsec

WPI

Confidence Intervals for Regression Parameters (1 of 2)

- Since parameters a and b are based on measured values with error, the predicted value (y) is also subject to errors
- Can derive confidence intervals for a and b
- First, need estimate of variance of a and b

  $s^2 = SSE / (n-2)$
  - With n measurements and two variables, the degrees of freedom are n-2
- Expand SSE

  $= \Sigma e_i^2 = \Sigma(y_i - a - bx_i)^2 = \Sigma[(y_i - \underline{y}) - b(x_i - \underline{x})]^2$

WPI

Confidence Intervals for Regression Parameters (2 of 3)

- Helpful to represent SSE as:

  $SSE = S_{yy} - 2bS_{xy} + b^2 2S_{xx} = S_{yy} - bS_{xy}$
- Where

  $S_{xx} = \Sigma(x_i - \underline{x})^2 = \Sigma x_i^2 - (\Sigma x_i)^2 / n$

  $S_{yy} = \Sigma(y_i - \underline{y})^2 = \Sigma y_i^2 - (\Sigma y_i)^2 / n$

  $S_{xy} = \Sigma(x_i - \underline{x})(y_i - \underline{y}) = \Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i) / n$
- So, $s^2 = SSE / (n-2)$

  $= S_{yy} - bS_{xy} / (n-2)$

WPI

Confidence Intervals for Regression Parameters (3 of 3)

- Conf interval for slope (b) and y intercept (a):

  $[b_1,b_2] = b \pm t_{[1-\alpha/2;n-2]} s / \mathrm{sqrt}(S_{xx})$

  $[a_1,a_2] = a \pm t_{[1-\alpha/2;n-2]} \dfrac{s \times \mathrm{sqrt}(\Sigma x_i^2)}{\mathrm{sqrt}(nS_{xx})}$

- Finally, for prediction y_p can determine interval $[y_{p1}, y_{p2}]$:

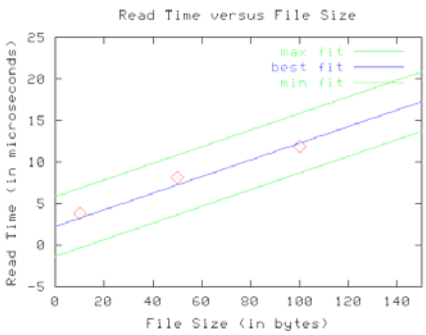
  $= y_p \pm t_{[1-\alpha/2;n-2]} s \times \mathrm{sqrt}\,(1 + 1/n + (x_p-\underline{x})^2/S_{xx})$

Regression Conf Interval Example (1 of 2)

$y = 2.24 + 0.1002x$

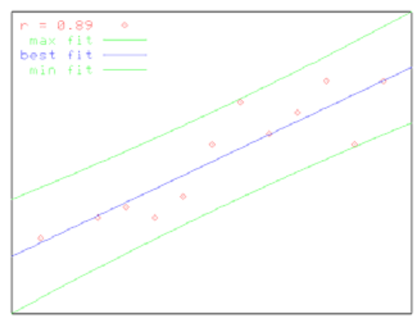
- $\Sigma x_i = 16{,}660.0$
- $\Sigma y_i = 1685.3$
- $\Sigma x_i y_i = 12{,}691{,}033.0$
- $\Sigma x_i^2 = 126{,}262{,}600.0$
- $\underline{x} = 2380$
- $\underline{y} = 240.76$
- $b = \dfrac{(7)(12691033) - (16660)(1685.3)}{(7)(126262600) - (16660)^2}$
- $a = 240.76 - .1002(2380)$
  $= 2.24$
- $y = 2.24 + 0.1002x$
- $S_{xx} = 126262600 - 16660^2/7$
  $= 86{,}611{,}800$
- $S_{yy} = 1275670.43 - (1685.3)^2 / 7$
  $= 869{,}922.42$
- $S_{xy} = 12691033-(16660)(1685.3)/7$
  $= 8{,}680{,}019$
- $s^2 = \dfrac{869922.42 - 0.1002(8680019)}{(7-2)}$
- Std dev $s = \mathrm{sqrt}(36.9027) = 6.0748$
- 90% conf interval
  - $[b_1,b_2] = [0.099, 0.102]$
  - $[a_1,a_2] = [-3.35, 7.83]$

Regression Conf Interval Example (2 of 2)

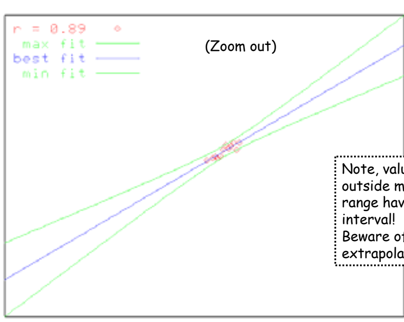


(Zoom)

Another Regression Conf Interval Example (1 of 2)



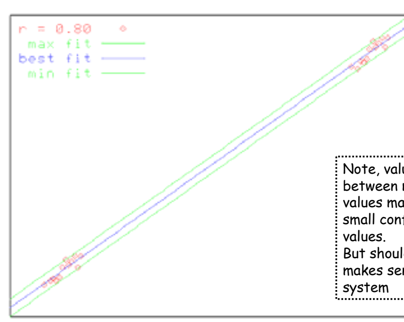
Another Regression Conf Interval Example (2 of 2)



(Zoom out)

Note, values outside measured range have larger interval!
Beware of large extrapolations

Another Regression Conf Interval Example



Note, values between measured values may have small confidence values.
But should verify makes sense for system

Correlation

- After developing regression model, useful to know how well the regression equation fits the data
  - *Coefficient of determination*
    - Determines how much of the total variation is explained by the linear model
  - *Correlation coefficient*
    - Square root of the coefficient of determination

Coefficient of Determination

- Earlier: $SSE = S_{yy} - bS_{xy}$
- Let: $SST = S_{yy}$ and $SSR = bS_{xy}$
- Now: $SST = SSR + SSE$
  - Total variation (SST) has two components
    - SSR portion explained by regression
    - SSE is model error (distance from line)
- Fraction of total variation explained by model line:

  $r^2 = SSR / SST = (SST - SSE) / SST$
  - Called *coefficient of determination*
- How "good" is the regression model? Roughly:
  - $0.8 <= r^2 <= 1$ strong
  - $0.5 <= r^2 < 0.8$ medium
  - $0 <= r^2 < 0.5$ weak

Correlation Coefficient

- Square root of coefficient of determination is the correlation coefficient. Or:

  $$r = S_{xy} / \mathrm{sqrt}(S_{xx}S_{yy})$$
- Note, equivalently:

  $$r = b\ \mathrm{sqrt}(S_{xx}/S_{yy}) = \mathrm{sqrt}(SSR/SST)$$
  - Where $b = S_{xy}/S_{xx}$ is slope of regression model line
- Value of r ranges between -1 and +1
  - +1 is perfect linear positive relationship
    - Change in x provides corresponding change in y
  - -1 is perfect linear negative relationship

Correlation Example

- From Read Size vs. Time model, correlation:

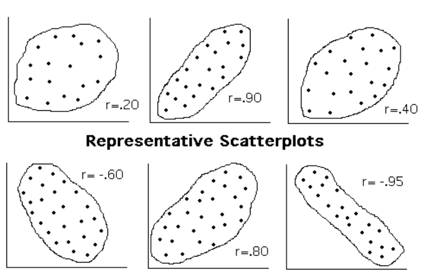
  $r = b\ \mathrm{sqrt}(S_{xx}/S_{yy})$

  $= 0.1002\ \mathrm{sqrt}(86{,}611{,}800 / 869{,}922.4171)$

  $= 0.9998$
- Coefficient of determination:

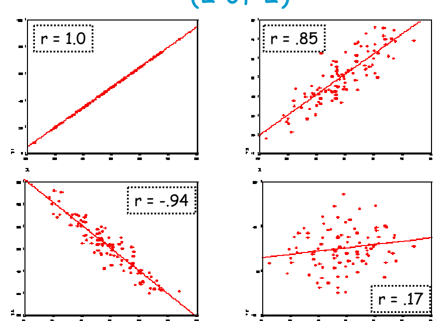
  $r^2 = (0.9998)^2 = 0.9996$
- So, 99.96% of the variation in time to read a file is explained by the linear model
- Note, correlation is not causation!
  - Large file maybe does *cause* more time to read
  - But, for example, time of day does not *cause* message to take longer

Correlation Visual Examples (1 of 2)



(http://peace.saumag.edu/faculty/Kardas/Courses/Statistics/Lectures/C4CorrelationReg.html)

Correlation Visual Examples (2 of 2)



(http://www.psychstat.smsu.edu/introbook/SBK17.htm)

Multiple Linear Regression (1 of 2)

- Include effects of several input variables that are linearly related to one output
- Straight-forward extension of single regression
- First, consider two variables. Need:

$$y = b_0 + b_1x_1 + b_2x_2$$

- Make n measurements of (x_{1i}, x_{2i}, y_i) and:

$$y_i = b_0 + b_1x_{1i} + b_2x_{2i} + e_i$$

- As before, want to minimize sum square of residual errors (the e_i's):

$$SSE = \Sigma e_i^2 = \Sigma(y_i-b_0-b_1x_{1i}-b_2x_{2i})^2$$

Multiple Linear Regression (2 of 2)

- As before, minimal when partial derivatives 0

$$nb_0 + b_1\Sigma x_{1i} + b_2\Sigma x_{2i} = \Sigma y_i$$
$$b_0\Sigma x_{1i} + b_1\Sigma x_{1i}^2 + b_2\Sigma x_{1i}x_{2i} = \Sigma x_{1i}y_i$$
$$b_0\Sigma x_{2i} + b_1\Sigma x_{1i}x_{2i} + b_2\Sigma x_{2i}^2 = \Sigma x_{2i}y_i$$

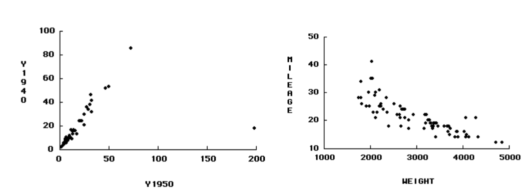
- Three equations in three unknowns (b_0, b_1, b_2)
  - Solve using wide variety of software
- Generalize:

$$y = b_0 + b_1x_1 + \ldots + b_kx_k$$

- Can represent equations as matrix and solve using available software

Verifying Linearity (1 of 2)

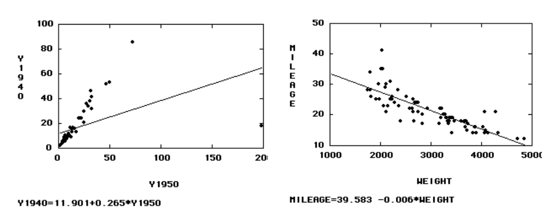
- Should do by visual check before regression



(http://peace.saumag.edu/faculty/Kardas/Courses/Statistics/Lectures/C4CorrelationReg.html)

Verifying Linearity (2 of 2)

- Linear regression may not be best model



(http://peace.saumag.edu/faculty/Kardas/Courses/Statistics/Lectures/C4CorrelationReg.html)

Outline

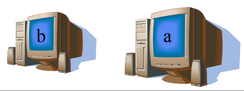
- Introduction
- Basics
- Indices of Central Tendency
- Indices of Dispersion
- Comparing Systems
- Misc
- Regression
- ANOVA

Analysis of Variance (ANOVA)

- Partitioning variation into part that can be explained and part that cannot be explained
- Example:
  - Easy to see regression that explains 70% of variation is not as good as one that explains 90% of variation
  - But how much of the explained variation is good?
- Enter: ANOVA

(Prof. David Lilja, ECE Dept., University of Minnesota)

Before-and-After Comparison



Measurement (i)	Before (b_i)	After (a_i)	Difference ($d_i = b_i - a_i$)
1	85	86	-1
2	83	88	-5
3	94	90	4
4	90	95	-5
5	88	91	-3
6	87	83	4

Mean of differences $\underline{d}$ = -1, Standard deviation s_d = 4.15

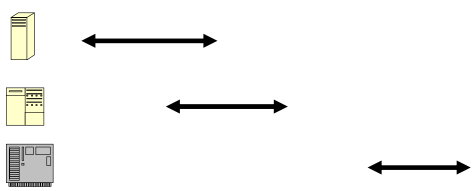
Before-and-After Comparison

Mean of differences $\underline{d}$ = -1
Standard deviation s_d = 4.15

- From mean of differences, appears that system change reduced performance
- However, standard deviation is large
- Is the variation between the two systems (alternatives) greater than the variation (error) in the measurements?
- Confidence intervals can work, but what if there are more than two alternatives?

Comparing More Than Two Alternatives

- Naïve approach
  - Compare confidence intervals



- Need to do for all pairs. Grows quickly.
- Ex- 7 alternatives would require 21 pair-wise comparisons [(7 choose 2) = (7)(6) / (2)(1) = 42]
- Plus, would not be surprised to find 1 pair differed (at 95%)

ANOVA – Analysis of Variance (1 of 2)

- Separates total variation observed in a set of measurements into:
  - (1) Variation within one system
    - Due to uncontrolled measurement errors
  - (2) Variation between systems
    - Due to real differences + random error
- Is variation (2) statistically greater than variation (1)?

ANOVA – Analysis of Variance (2 of 2)

- Make n measurements of k alternatives
- y_{ij} = ith measurement on jth alternative
- Assumes errors are:
  - Independent
  - Normally distributed

(Long example next)

All Measurements for All Alternatives

	Alternatives					
Measure-ments	1	2	…	j	…	k
1	y_{11}	y_{12}	…	y_{1j}	…	y_{k1}
2	y_{21}	y_{22}	…	y_{2j}	…	y_{2k}
…	…	…	…	…	…	…
i	y_{i1}	y_{i2}	…	y_{ij}	…	y_{ik}
…	…	…	…	…	…	…
n	y_{n1}	y_{n2}	…	y_{nj}	…	y_{nk}
Column mean	$y_{.1}$	$y_{.2}$	…	$y_{.j}$	…	$y_{.k}$
Effect	α_1	α_2	…	α_j	…	α_k

Column Means

- Column means are average values of all measurements within a single alternative
  - Average performance of one alternative

$$\bar{y}_{.j} = \frac{\sum_{i=1}^{n} y_{ij}}{n}$$

	Alternatives					
Measure-ments	1	2	...	*j*	...	*k*
1	y_{11}	y_{12}	...	y_{1j}	...	y_{k1}
2	y_{21}	y_{22}	...	y_{2j}	...	y_{2k}
...	...	...	...	...	...	...
i	y_{i1}	y_{i2}	...	y_{ij}	...	y_{ik}
...	...	...	...	...	...	...
n	y_{n1}	y_{n2}	...	y_{nj}	...	y_{nk}
Column mean	$y_{.1}$	$y_{.2}$	...	$y_{.j}$	...	$y_{.k}$
Effect	α_1	α_2	...	α_j	...	α_k

Error = Deviation From Column Mean

- $y_{ij} = y_{.j} + e_{ij}$
- Where e_{ij} = error in measurements

	Alternatives					
Measure-ments	1	2	...	*j*	...	*k*
1	y_{11}	y_{12}	...	y_{1j}	...	y_{k1}
2	y_{21}	y_{22}	...	y_{2j}	...	y_{2k}
...	...	...	...	...	...	...
i	y_{i1}	y_{i2}	...	y_{ij}	...	y_{ik}
...	...	...	...	...	...	...
n	y_{n1}	y_{n2}	...	y_{nj}	...	y_{nk}
Column mean	$y_{.1}$	$y_{.2}$	...	$y_{.j}$	...	$y_{.k}$
Effect	α_1	α_2	...	α_j	...	α_k

Overall Mean

- Average of all measurements made of all alternatives

$$\bar{y}_{..} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n} y_{ij}}{kn}$$

	Alternatives					
Measure-ments	1	2	...	*j*	...	*k*
1	y_{11}	y_{12}	...	y_{1j}	...	y_{k1}
2	y_{21}	y_{22}	...	y_{2j}	...	y_{2k}
...	...	...	...	...	...	...
i	y_{i1}	y_{i2}	...	y_{ij}	...	y_{ik}
...	...	...	...	...	...	...
n	y_{n1}	y_{n2}	...	y_{nj}	...	y_{nk}
Column mean	$y_{.1}$	$y_{.2}$	...	$y_{.j}$	...	$y_{.k}$
Effect	α_1	α_2	...	α_j	...	α_k

Effect = Deviation From Overall Mean

- $y_{.j} = y_{..} + \alpha_j$
- α_j = deviation of column mean from overall mean
  = effect of alternative j

	Alternatives					
Measure-ments	1	2	...	*j*	...	*k*
1	y_{11}	y_{12}	...	y_{1j}	...	y_{k1}
2	y_{21}	y_{22}	...	y_{2j}	...	y_{2k}
...	...	...	...	...	...	...
i	y_{i1}	y_{i2}	...	y_{ij}	...	y_{ik}
...	...	...	...	...	...	...
n	y_{n1}	y_{n2}	...	y_{nj}	...	y_{nk}
Col mean	$y_{.1}$	$y_{.2}$	...	$y_{.j}$	...	$y_{.k}$
Effect	α_1	α_2	...	α_j	...	α_k

Effects and Errors

- *Effect* is distance from overall mean
  - Horizontally across alternatives
- *Error* is distance from column mean
  - Vertically within one alternative
  - Error across alternatives, too
- Individual measurements are then:

$$y_{ij} = \bar{y}_{..} + \alpha_j + e_{ij}$$

Sum of Squares of Differences

- ***SST*** = differences between each measurement and overall mean
- ***SSA*** = variation due to effects of alternatives
- ***SSE*** = variation due to errors in measurements

$$SSA = n\sum_{j=1}^{k}\left(\bar{y}_{.j} - \bar{y}_{..}\right)^2$$

$$SSE = \sum_{j=1}^{k}\sum_{i=1}^{n}\left(y_{ij} - \bar{y}_{.j}\right)^2$$

$$SST = \sum_{j=1}^{k}\sum_{i=1}^{n}\left(y_{ij} - \bar{y}_{..}\right)^2$$

$$\boxed{SST = SSA + SSE}$$

ANOVA

- Separates variation in measured values into:
  1. Variation due to effects of alternatives
     - SSA – variation across columns
  2. Variation due to errors
     - SSE – variation within a single column
- If differences among alternatives are due to real differences:

→ SSA statistically greater than SSE

Comparing SSE and SSA

- Simple approach
  - *SSA* / *SST* = fraction of total variation explained by differences among alternatives
  - *SSE* / *SST* = fraction of total variation due to experimental error
- But is it statistically significant?
- Variance = mean square values

  = total variation / degrees of freedom

  s_x^2 = SSx / df(SSx)
- (Degrees of freedom are number of independent terms in sum)

Degrees of Freedom for Effects

- *df*(*SSA*) = *k* – 1, since *k* alternatives

	Alternatives					
Measure-ments	1	2	...	*j*	...	*k*
1	y_{11}	y_{12}	...	y_{1j}	...	y_{k1}
2	y_{21}	y_{22}	...	y_{2j}	...	y_{2k}
...	...	...	...	...	...	...
i	y_{i1}	y_{i2}	...	y_{ij}	...	y_{ik}
...	...	...	...	...	...	...
n	y_{n1}	y_{n2}	...	y_{nj}	...	y_{nk}
Column mean	$y_{.1}$	$y_{.2}$	...	$y_{.j}$	...	$y_{.k}$
Effect	α_1	α_2	...	α_j	...	α_k

Degrees of Freedom for Errors

- *df*(*SSE*) = *k*(*n* – 1), since *k* alternatives, each with (*n* – 1) *df*

	Alternatives					
Measure-ments	1	2	...	*j*	...	*k*
1	y_{11}	y_{12}	...	y_{1j}	...	y_{k1}
2	y_{21}	y_{22}	...	y_{2j}	...	y_{2k}
...	...	...	...	...	...	...
i	y_{i1}	y_{i2}	...	y_{ij}	...	y_{ik}
...	...	...	...	...	...	...
n	y_{n1}	y_{n2}	...	y_{nj}	...	y_{nk}
Column mean	$y_{.1}$	$y_{.2}$	...	$y_{.j}$	...	$y_{.k}$
Effect	α_1	α_2	...	α_j	...	α_k

Degrees of Freedom for Total

- *df*(*SST*) = *df*(*SSA*) + *df*(*SSE*) = *kn* - 1

	Alternatives					
Measure-ments	1	2	...	*j*	...	*k*
1	y_{11}	y_{12}	...	y_{1j}	...	y_{k1}
2	y_{21}	y_{22}	...	y_{2j}	...	y_{2k}
...	...	...	...	...	...	...
i	y_{i1}	y_{i2}	...	y_{ij}	...	y_{ik}
...	...	...	...	...	...	...
n	y_{n1}	y_{n2}	...	y_{nj}	...	y_{nk}
Column mean	$y_{.1}$	$y_{.2}$	...	$y_{.j}$	...	$y_{.k}$
Effect	α_1	α_2	...	α_j	...	α_k

Variances from Sum of Squares (Mean Square Value)

$$s_a^2 = \frac{SSA}{k-1}$$

$$s_e^2 = \frac{SSE}{k(n-1)}$$

Comparing Variances

- Use F-test to compare ratio of variances
  - An F-test is used to test if the standard deviations of two populations are equal.

$$F = \frac{s_a^2}{s_e^2}$$

$$F_{[1-\alpha;df(num),df(denom)]} = \text{tabulated critical values}$$

- If $F_{computed} > F_{table}$ for a given α
  → We have $(1 - \alpha)$ * 100% confidence that variation due to actual differences in alternatives, SSA, is statistically greater than variation due to errors, SSE.

ANOVA Summary

Variation	Alternatives	Error	Total
Sum of squares	SSA	SSE	SST
Deg freedom	$k-1$	$k(n-1)$	$kn-1$
Mean square	$s_a^2 = SSA/(k-1)$	$s_e^2 = SSE/[k(n-1)]$	
Computed F	s_a^2/s_e^2		
Tabulated F	$F_{[1-\alpha;(k-1),k(n-1)]}$		

(Example next)

ANOVA Example (1 of 2)

	Alternatives			
Measurements	**1**	**2**	**3**	**Overall mean**
1	0.0972	0.1382	0.7966	
2	0.0971	0.1432	0.5300	
3	0.0969	0.1382	0.5152	
4	0.1954	0.1730	0.6675	
5	0.0974	0.1383	0.5298	
Column mean	0.1168	0.1462	0.6078	0.2903
Effects	-0.1735	-0.1441	0.3175	

ANOVA Example (2 of 2)

Variation	Alternatives	Error	Total
Sum of squares	$SSA = 0.7585$	$SSE = 0.0685$	$SST = 0.8270$
Deg freedom	$k-1=2$	$k(n-1)=12$	$kn-1=14$
Mean square	$s_a^2 = 0.3793$	$s_e^2 = 0.0057$	
Computed F	$0.3793/0.0057 = 66.4$		
Tabulated F	$F_{[0.95;2,12]} = 3.89$		

- SSA/SST = 0.7585/0.8270 = 0.917
  → 91.7% of total variation in measurements is due to differences among alternatives
- SSE/SST = 0.0685/0.8270 = 0.083
  → 8.3% of total variation in measurements is due to noise in measurements
- Computed *F* statistic > tabulated *F* statistic
  → 95% confidence that differences among alternatives are statistically significant.

ANOVA Summary

- Useful for partitioning total variation into components
  - Experimental error
  - Variation among alternatives
- Compare more than two alternatives
- Note, does not tell you *where* differences may lie
  - Use confidence intervals for pairs
  - Or use contrasts