

When is Enough Data, Enough?



*Midwest Region University
Transportation Center,
UW-Madison Campus*

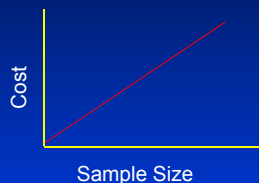
Robert L. Schmitt, Univ. of Wis.- Platteville
Sam Owusu-Ababio, Univ. of Wis.- Platteville
Richard M. Weed, formerly NJDOT
Erik V. Nordheim, Univ. of Wis.- Madison

When is Enough Data, Enough?

Multiple Choice:

- a. *There never is enough data*
- b. *There is too much data*
- c. *It all depends*
- d. *All the above*
- e. *None of the above*

Determining the amount of data



Audience

1. Technical
2. Administrative
3. Legislative

Larger sample sizes yield smaller error.
Requires greater resources and cost.

Determining sample size

Balance between statistics and cost

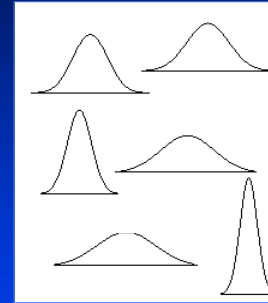
1. What can you afford
2. Rule of thumb
 - historical precedence
 - past experience
 - some consideration of sample error
3. Make up of sub-groups (Functional Class, Regions)
 - What do you hope to understand between sub-groups
4. Statistical determination

Determining Sample Size

Three key pieces of information required:

1. An estimate of the population Standard Deviation
2. A desired Level of Precision or Confidence that the Sample Result will fall within a certain range (result +/- sampling error) of true population values
3. An acceptable Level of Sampling Error

Standard Deviation



Continuous Data:

- Number of obstructed drains
- Length of blocked ditch
- L.O.S.

Discrete Data:

- Operability of vending machines
- Proportion of vegetation obstruction

Continuous data and variability



Guardrail

Total linear feet of guardrail, per 0.10 mile section.



Total linear feet of defective guardrail, per 0.10 mile section.

$$S = \sqrt{\sum(X_i - \bar{X})^2 / n - 1}$$

WSDOT

Hazardous Debris – Stratified by Sub-Groups

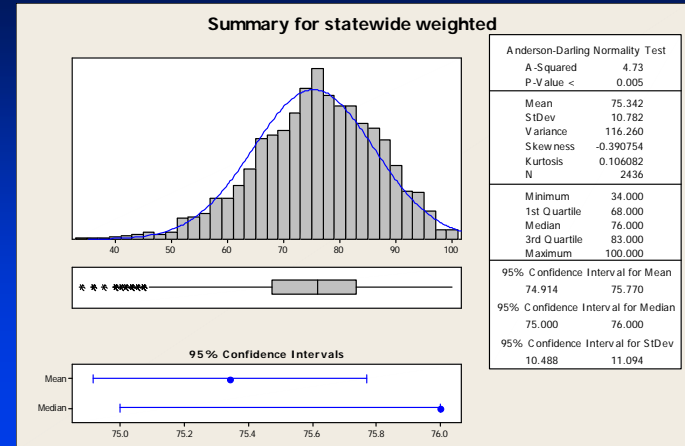
Attribute Division	n	Mean, count	Std Dev, count
Statewide	1892	0.25	1.041
Divided Hwy	345	0.53	1.58
Undivided Hwy	1547	0.18	0.77
District 1	231	1.02	2.21
District 2	239	0.13	0.46
District 3	239	0.13	0.47
County 1	14	0.21	0.42
County 2	19	0.10	0.31
County 3	29	0.00	0.00

Level of Service

$$LOS_s = \frac{\sum_{j=1}^N LOS_j W_j}{\sum_{j=1}^N W_j}$$

Element	Weight
Pavement	30
Traffic Control	20
Shoulders & Ditches	16
Roadside	15
Drainage	14
Environmental	5

Continuous data – L.O.S.



Discrete data and variability



Culvert Damage

Yes - Major Culvert damage on both ends count as one defect.

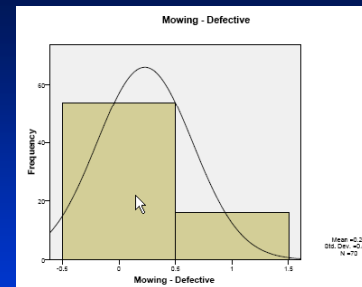
No - Minor Culvert damage is acceptable.



$$\sigma_p = \sqrt{p(1-p)/n}$$

Iowa DOT

Discrete data and variability



Acceptable = $53/70 = 0.76$

Unacceptable = $17/70 = 0.24$

$$\sigma_p = \sqrt{0.76(1-0.76)/70}$$

$$\sigma_p = 0.05$$

What if defective mowing is 50/50 Pass/Fail ?

$$\sigma_p = \sqrt{0.50(1-0.50)/70}$$

$$\sigma_p = 0.06$$

Standard deviation increases 20%.
Measuring variability is very important.

Determining Sample Size

Three key pieces of information required:

1. An estimate of the population Standard Deviation
2. A desired Level of Precision or Confidence that the Sample Result will fall within a certain range (result +/- sampling error) of true population values
3. An acceptable Level of Sampling Error

Accuracy, Repeatability, and Reproducibility

Good Precision
Poor Accuracy

Poor Precision
Good Accuracy

Good Precision
Good Accuracy



Sample Size and Confidence Interval

$$C.I. = \overline{LOS}_s \pm \left(Z * \frac{s}{\sqrt{n}} \right)$$

C.I. = +/- 1 LOS

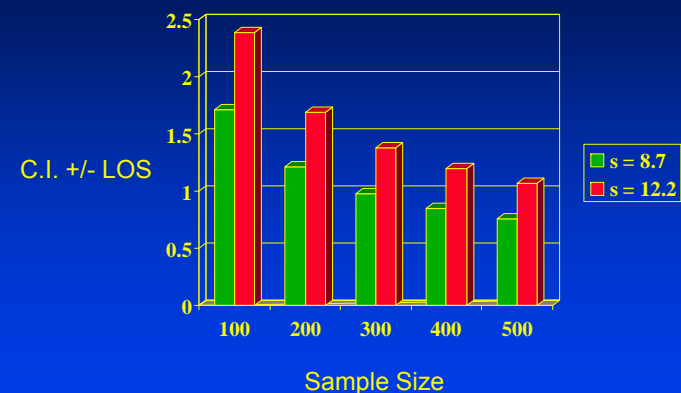
Z = 1.96 (95% probability)

s = Standard Deviation

In 2002, s = 12.2 LOS

In 2004, s = 8.7 LOS

Confidence Interval and Sample Size



Determining Sample Size

Three key pieces of information required:

1. An estimate of the population Standard Deviation
2. A desired Level of Precision or Confidence that the Sample Result will fall within a certain range (result +/- sampling error) of true population values
3. *An acceptable Level of Sampling Error*

Errors and Risks in Sample Confidence

“Probability” we can take results as “accurate representation” of entire roadway condition.

Typically a 95% Probability is used

19 times out of 20 we would expect results in this range or Confidence Interval

Risks

Suppose (hypothesize):

H_0 : Mean difference between data collectors is zero

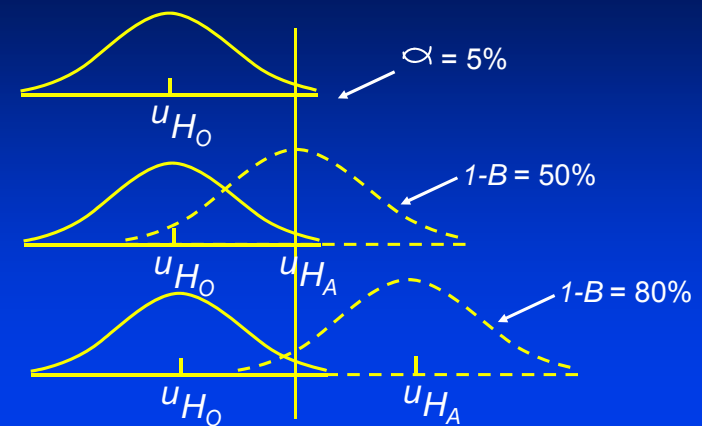
H_A : Mean diff. between data collectors is not zero

Risks:

Type I = Reject H_0 when H_0 is true: $\alpha = 5\%$

Type II = Accept H_0 when H_0 is false: $B = 20\%$

Illustration of Risks



Difference of Two Data Collectors



$$n = \left[\frac{(Z_{\alpha/2=0.025} + Z_{\beta=0.1}) * s}{d} \right]^2$$



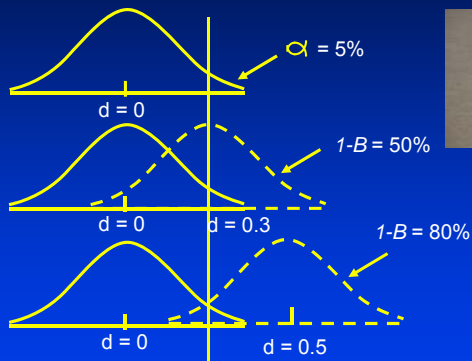
- α Error for detecting no difference
- β Error for detecting true mean diff.
- $1-\beta$ Power

Variability, s



Using same comparison location will reduce variability and sample size

Difference of Two Data Collectors



Pieces of Hazardous Debris

Difference of Two Data Collectors

Std Dev, count	Allowable Difference, count			
	1	2	3	4
0.10	0.3	0.2	0.2	0.1
0.20	0.6	0.4	0.3	0.3
0.30	0.8	0.6	0.5	0.4

Recommended Approach for Verification Procedures

- Incorporate actual field variation into procedure
- Assess risks of procedure
- Compute allowable mean difference of data collectors

Are years different or not
with risk = 5% ?

$$\left| \bar{X}_1 - \bar{X}_2 \right| < t_{\alpha, y} \sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)}$$

Are years different or not ?

Interstate Element	2002 Mean (250)	2004 Mean (265)	2002 S.D.	2004 S.D.	Pooled S.D.	Diff.?
Shoulders	80	87	17.2	15.7	1.45	Yes
Drainage	78	70	22.4	28.1	2.33	Yes
Roadside	74	76	14.9	13.9	1.27	No
Traffic	76	82	23.0	20.0	1.90	Yes
Entire System	78	80	11.3	9.3	0.92	No

When is Enough Data, Enough?

Multiple Choice:

- There never is enough data*
- There is too much data*
- It all depends*
- All the above*
- None of the above*

Summary – when determining sample size

1. Measure and estimate population Standard Deviation
2. Define a desired Level of Precision or Confidence that the Sample Result will fall within a certain range (result +/- sampling error) of true population values
3. Assess an acceptable Level of Sampling Error or Risk
4. Cost

Acknowledgments

*Midwest Region University
Transportation Center,
UW-Madison Campus*

www.mrutc.org/research/0604

Agencies

California	Texas
Florida	Utah
Indiana	Virginia
New York	Washington
North Carolina	Wisconsin