

LEAN Ohio
GREEN BELT
Transforming the Public Sector
Normal Data

SIMPLER. FASTER. BETTER. LESS COSTLY. lean.ohio.gov

LEAN Ohio

Making state government in Ohio
simpler, faster, better,
and less costly.

SIMPLER. FASTER. BETTER. LESS COSTLY. lean.ohio.gov

Determining Normality

- Not all data is normal
 - Belts must know whether or not the data is normal as different tests apply in different circumstances
- One of the first steps is to determine normality
- Normal data is defined as data that has “normal” variation

SIMPLER. FASTER. BETTER. LESS COSTLY. lean.ohio.gov

Probability Distributions

Normal Data – Why should I care!

- Normally distributed data exhibit predictable traits and probabilities
- In practice, we are frequently confronted with data that is not normal.
- The first step to take is to look at how the data is distributed

SIMPLER. FASTER. BETTER. LESS COSTLY.

lean.ohio.gov

Distributions

When we measure a quantity in a large number of individuals we call the pattern of values obtained a distribution.

SIMPLER. FASTER. BETTER. LESS COSTLY.

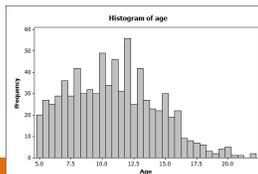
lean.ohio.gov

What is a distribution?

A way of describing and viewing the data that you have

– Focuses on:

- Shape of the data
- 'Range' of the data e.g. maximum and minimum point



SIMPLER. FASTER. BETTER. LESS COSTLY.

Distributions

- Normal distribution – The granddaddy of them all!
 - Also known as the Gaussian distribution (after Gauss, German mathematician)
 - e.g. heights of adult men in the UK
- T-distribution
 - Similar shape to Normal, but is more spread out with longer tails
 - Useful for calculating confidence intervals
- Chi-squared distribution
 - Right-skewed distribution
 - Useful for analysing categorical data
- F-distribution
 - Skewed to the right
 - Useful for comparing variances and more than 2 means (i.e. > 2 groups)
- Binomial distribution
 - Could be skewed to the right or left (!)
 - Good for analysing proportion data – i.e. it is either one thing, or another, such as an animal either has a disease or does not have a disease

Our focus today

SIMPLER. FASTER. BETTER. LESS COSTLY. lean.ohio.gov

Distributions (cont.)

- Poisson distribution
 - Right skewed
 - Good for analysing count data – i.e. the number of hospital admissions per day, the number of parasitic eggs per gram of faecal sample
- Many of these distributions approximate normal when your sample size increases

SIMPLER. FASTER. BETTER. LESS COSTLY. lean.ohio.gov

Normal Data



- In this context the name “normal” causes much confusion. In statistics it is just a name
- Indeed, in some arenas normal distributions are rare
- Various methods of analysis make assumptions about normality, including:
 - correlation, regression, t tests, and analysis of variance

SIMPLER. FASTER. BETTER. LESS COSTLY. lean.ohio.gov

Normal Data

- When data do not have a normal distribution we can either transform the data (for example, by taking logarithms) or use a method that does not require the data to be normally distributed.
- LeanOhio avoids transforming data sets

SIMPLER. FASTER. BETTER. LESS COSTLY.

lean.ohio.gov

Normal distribution

- In relation to **numerical data**
- A starting point – is your data normal?
- Most commonly talked about
 - If your data is normal, a whole range of tests you can apply
 - There are others to use if your data is not normal, so don't worry if it isn't!

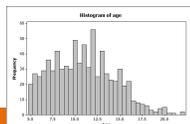


SIMPLER. FASTER. BETTER. LESS COSTLY.

lean.ohio.gov

4 steps to Normality!

- Plot your data
 - Create a histogram with frequencies and determine by eye
 - Does it look bell-shaped and symmetrical?
 - Does it look unimodal i.e. does it only have one peak?
 - Subjective measurement, but you should be doing this anyway!



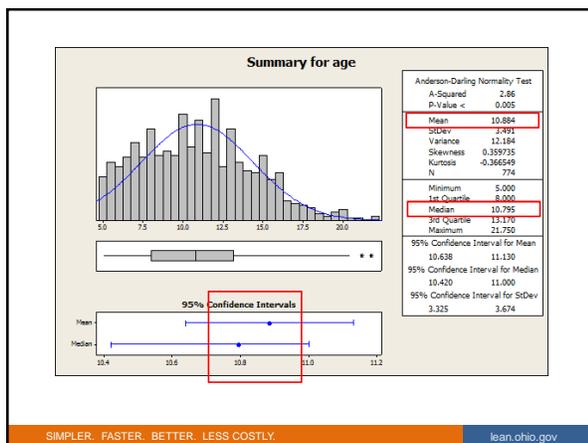
SIMPLER. FASTER. BETTER. LESS COSTLY.

4 steps (continued)

- How different are the mean and median?
 - Mean = Total of your data added up/total no. of measurements
 - Median = The midpoint of your values i.e. what is the 'halfway' value in your data?
 - If they are very different, the data is probably not normally distributed
 - If they are very similar, your data could be normally distributed
 - Another rule of thumb, so not always correct!

SIMPLER. FASTER. BETTER. LESS COSTLY.

lean.ohio.gov

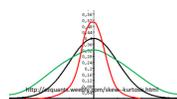


SIMPLER. FASTER. BETTER. LESS COSTLY.

lean.ohio.gov

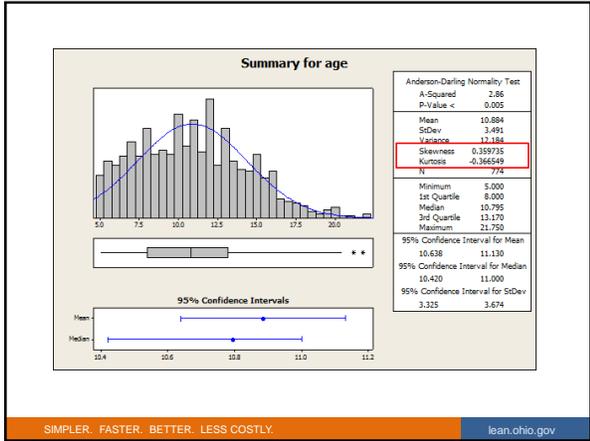
4 steps (continued)

- Skewness and kurtosis
 - Skewness (how symmetrical the data is)
 - Normal – this value is 0
 - Right-skewed distribution – positive value
 - Left-skewed distribution – negative value
 - Kurtosis (the 'peakedness' of the data) – does your data have a pointy bit, or is it flat?
 - Normal – this value is 0
 - Sharply peaked data – positive value
 - Flat peaked data – negative value



SIMPLER. FASTER. BETTER. LESS COSTLY.

lean.ohio.gov



4 steps (continued)

- Tests for normality
 - Shapiro-Wilk test (Ryan-Joiner test)
 - Kolmogorov-Smirnov test
 - **Anderson-Darling test**
 - Watch interpretation of p-values – if it is <0.05, it is not normal (reject null hypothesis of normality)
- The good news!
 - Computers do this for us so we don't have to!

SIMPLER. FASTER. BETTER. LESS COSTLY. lean.ohio.gov

4 Steps: Hypothesis Tests

- Anderson-Darling test is a hypothesis test
- It is the hypothesis test we run before we run our other hypothesis tests
- It tells us which test to run: normal or non-normal

SIMPLER. FASTER. BETTER. LESS COSTLY. lean.ohio.gov

Questions?

SIMPLER. FASTER. BETTER. LESS COSTLY.

lean.ohio.gov
