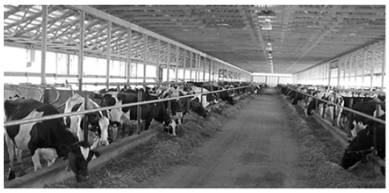


Data: Collection and Analysis

Daryl Nydam, DVM, PhD
Dept. of Population Medicine & Diagnostic Science
Cornell University, College of Vet Med
dvn2@cornell.edu

"There are three kinds of lies:

lies, damned lies, and statistics."

(Benjamin Disraeli as quoted by Mark Twain)

Acknowledgements:
Lorin Warnick, DVM, PhD
Ynte Schukken, DVM, PhD

Data

- Garbage in, Garbage out
- Lots of data that could be put to good use on livestock operations
- Challenge:
  - Capturing it in a usable form
  - Making sense of it
  - Doing something about it

Data

Decisions are based on information

Statistics help turn data into information

Data

1. Qualitative Data:
   - Non-numerical, descriptive stuff
     - Big/small, high/low, lots/little
2. Quantitative Data:
   - Numerical stuff
     - Continuous
       - Described by: central tendency, spread of data
     - Discrete (Categorical)
       - Described by: counts, risk, and rates

Levels of Measurement

- Dichotomous
  - yes/no; 0/1
- Categorical
  - Nominal (but, change into whole numbers)
  - Ordinal

Discrete Data

- Count events
- Divide by appropriate population
  - Rate
  - Ratio
  - Proportion

Discrete Data

- Incidence
- Prevalence
- Define numerators and denominators
  - Prevalence: how much disease
    - Existing "cases"
  - Incidence: how disease is changing
    - New "cases"

Levels of Measurement

- Continuous
  - Can take on any value on the number line

Continuous Data

- Normal
  - Bell shaped
  - Mean~median~mode
- Skewed
  - Any shape
  - Mean does not represent center very well

Continuous Data

- Normal
  - Mean
    - Milk production
  - Median
    - SCC
  - Mode

Continuous Data

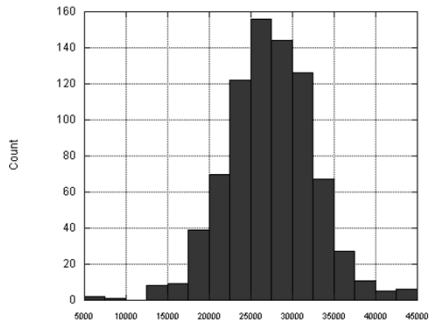
- Standard Deviation
  - Represents how far data values are from the mean (for **Normally** distributed data)
  - 1 SD = 65% of data
  - 2 SD = 95% of data
  - 3 SD = 99% of data

Continuous Data

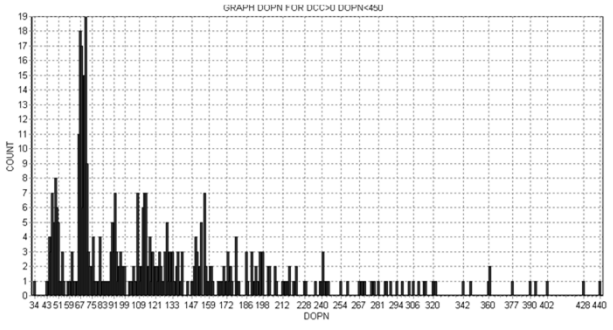
- Standard Error
  - The standard deviation of an average
  - ~ how confident one can be in the mean
  - = SD/ sqrt of N

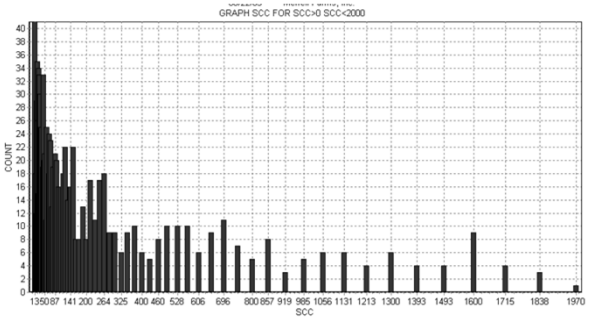
Distributions

- A function describing the frequency of measurements in a population
- Examples
  - ME305 milk production
  - Days Open
  - SCC



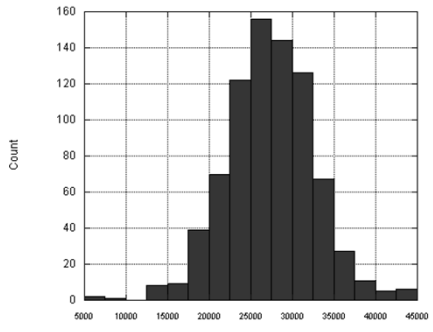
ME305





Distributions cont.

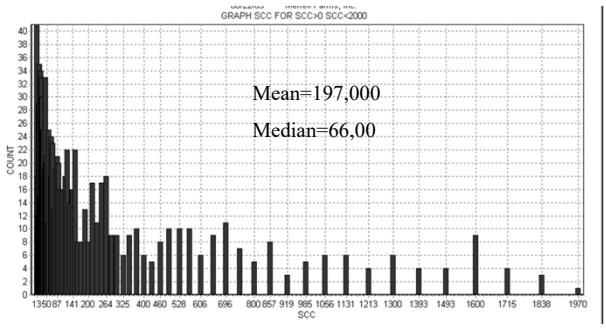
- Population parameters summarize information about a distribution
  - Mean or average
  - Standard deviation
  - Median
  - Mode
  - Minimum
  - Maximum



Mean = 27340

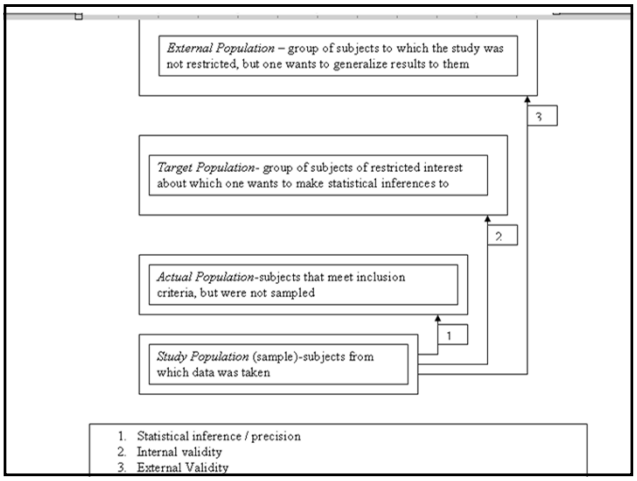
Median= 27330

ME305



Summary Approach to Analysis

- Level of measurement and distribution determine appropriate statistical method
- Sample statistics are calculated to estimate population parameters
- Make conclusions about population from sample data using hypothesis tests and/or confidence intervals

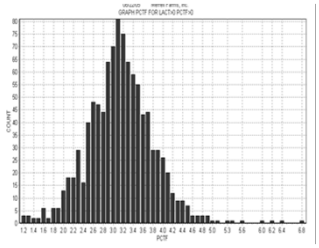


Using stats allows us to be cheap and lazy

- Can't get data on everything in a large population, so
- Take a SAMPLE
  - Not enough time to see every cow (lazy)
  - Not enough money to "test" every cow (cheap)
- Then say something about everything by only looking at part of it…
  - If you don't need to sample, you don't need statistics

Look at data before statistical analysis!

- Random sample
- Distribution
- Outliers and unreasonable values
- Bias?
  - Confounding
  - Selection
  - Classification
- Biological & medical basis for statistical comparisons



Steps for Data Analysis

1. Check distribution of data
   - Good time to find errors in data entry
2. Decide what analysis method to use
3. Evaluate how well assumptions are met
   - e.g. 2 independent samples, number of subjects, normality of data…
4. Interpret results in light of following:
   - Was method appropriate?
   - Is the result precise (or was sample size adequate)?
   - What are potential sources of bias?
     - Confounding, selection of sample, data recording etc.

Sampling

- Usually cannot take measurements on the whole population
- Rely on a representative *sample*
- Statistics are calculated for the sample as estimates of population parameters
  - Examples:
    - Sample mean
    - Sample standard deviation
- If a sample is random, the sample statistics often have known distributions

Sampling for inference

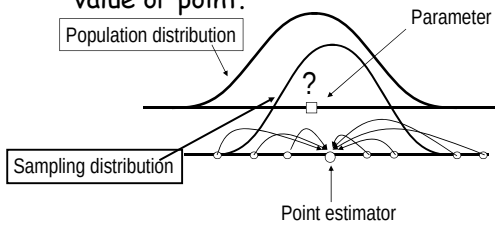
- Statistical inference is the process by which we acquire information about populations from samples.
- There are two types of inference:
  - Estimation
  - Hypotheses testing

Estimation

- The objective of estimation is to determine the value of a population parameter on the basis of a sample statistic.
- There are two types of estimators:
  Point Estimator
  Interval estimator

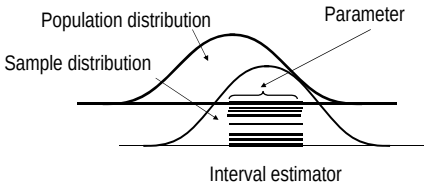
Point Estimator

A point estimator draws inference about a population by estimating the value of an unknown parameter using a single value or point.

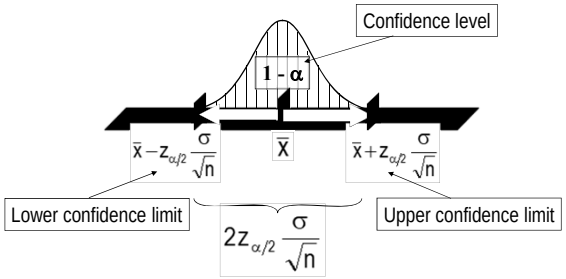


Interval Estimator

An interval estimator draws inferences about a population by estimating the value of an unknown parameter using an interval.



Confidence Interval for μ



The Confidence Interval for μ (σ is known)

- Four commonly used confidence levels

Confidence level	α	$\alpha/2$	$z_{\alpha/2}$
0.90	0.10	0.05	1.645
0.95	0.05	0.025	1.96
0.98	0.02	0.01	2.33
0.99	0.01	0.005	2.575

The Width of the Confidence Interval
(the precision of the estimate)

The width of the confidence interval is affected by

- the population standard deviation (σ)
- the confidence level (1-α)
- the sample size (n).

Confidence Interval

- 2000 milking cows, daily milk production 92 lbs, SD = 20
- 95% CI = 90 - 93
- 80% CI = 91 – 92.7

- 200 milking cows, daily milk production 92 lbs, SD=20
- 95% CI = 89 – 95

Fleiss Quadratic approximation

$$\text{Lower P} = \frac{(2np+z^2-1) - z\sqrt{z^2 - (2+1/n) + 4p(nq+1)}}{2(n+z^2)}$$

$$\text{Upper P} = \frac{(2np+z^2+1) + z\sqrt{z^2 + (2-1/n) + 4p(nq-1)}}{2(n+z^2)}$$

With
z : z value (1.96 for α=.05)
n : total observations
p : proportion with factor
q = 1 − p

This approximate confidence interval is valid even when p is near 0 or unity.

The Width of the Confidence Interval
(the precision of the estimate)

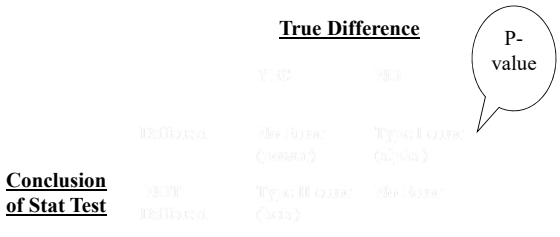
The width of the confidence interval for count data is affected by:

- the proportion, i.e. p and q
- the confidence level (1-a), i.e. Z value
- the sample size (n).
  - NOTE: there is no variance/SD in this one

Hypothesis Testing

- Question is stated in terms of a *null hypothesis*
  - For example: Propylene Glycol has no effect on occurrence of ketosis in cows compared to none
  - A=B
- Assuming the null hypothesis is true, what is the probability of the results observed?
- If the probability (p-value) is small, reject the null
  - Say A does not equal B
- Otherwise, fail to reject

Results of Stats v. the "Truth"



P-value Method

- The p-value provides information about the amount of statistical evidence that supports the alternative hypothesis.

- The p-value of a test is the probability of observing a test statistic at least as extreme as the one computed, given that the null hypothesis is true.

Probability of committing a Type 1 error

Conclusions of a Hypothesis Test

- If we reject the null hypothesis, we conclude that there is enough evidence to infer that the alternative hypothesis is true.
- If we do not reject the null hypothesis, we conclude that there is not enough statistical evidence to infer that the alternative hypothesis is true.

P-value interpretation

- P-value *does not* indicate the magnitude of the difference between groups, only the LIKELIHOOD that a difference of that magnitude could have arisen by chance alone

Results of Stats v. the "Truth"

True Difference

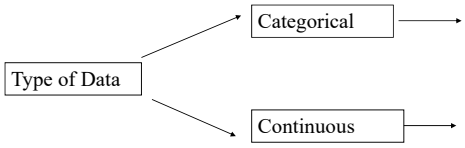
Conclusion of Stat Test

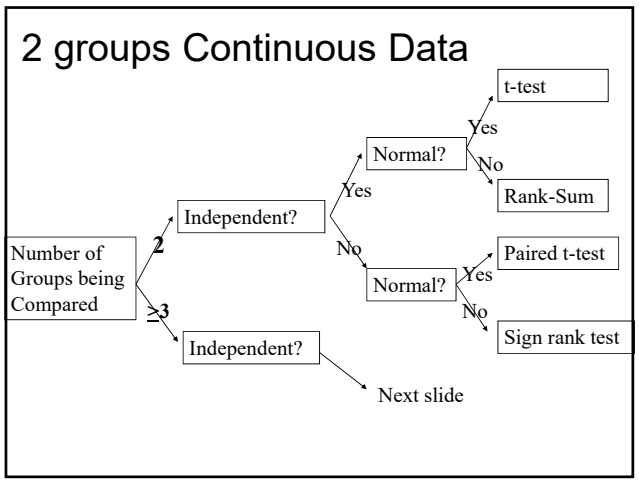
P-value

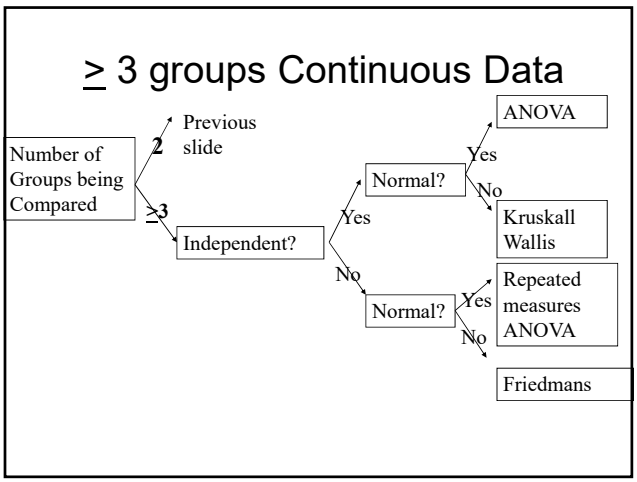
Type 1 error: say there is a difference when there really isn't one

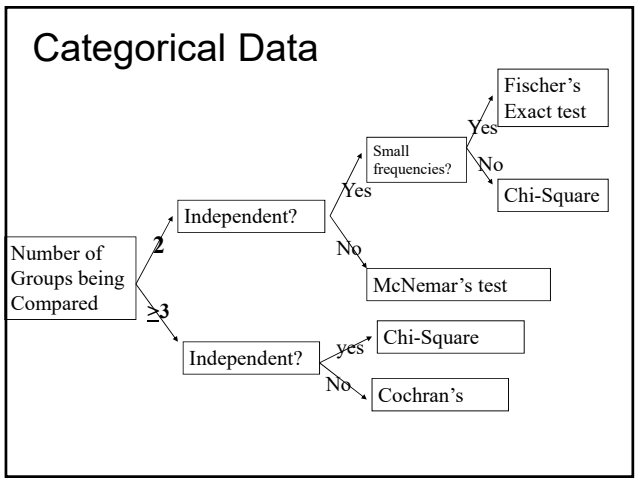
Type 2 error: say is NOT a difference when there really is one

Approach to Data




2 groups Continuous Data
Number of Groups being Compared
2
≥3
Independent?
Normal?
Yes
No
t-test
Rank-Sum
Paired t-test
Sign rank test
Independent?
Next slide


≥ 3 groups Continuous Data
Number of Groups being Compared
2
Previous slide
≥3
Independent?
Normal?
ANOVA
Kruskall Wallis
Repeated measures ANOVA
Friedmans


Categorical Data
Number of Groups being Compared
2
≥3
Independent?
Small frequencies?
Fischer's Exact test
Chi-Square
McNemar's test
Independent?
Chi-Square
Cochran's