Basic Biostatistics in Medical Research

Lecture 1: Basic Concepts

Mary J. Kwasny, ScD
Biostatistics Collaboration Center (BCC)
Department of Preventive Medicine
NU Feinberg School of Medicine
Objectives

• Assist participants in interpreting statistics published in medical literature

• Highlight different statistical methodology for investigators conducting own research

• Facilitate communication between medical investigators and biostatisticians
Lecture 1: Basic Concepts

Data Types
Summary Statistics
One Sample Inference
Data Types

• Categorical (Qualitative)
  – subjects sorted into categories
    • diabetic/non-diabetic
    • blood type: A/B/AB/O

• Numeric (Quantitative)
  – measurements or counts
    • weight in lbs
    • # of children
Categorical Data

• Nominal
  – categories unordered
    • blood type: A/B/AB/O
  – called binary or dichotomous if 2 categories
    • gender: M/F

• Ordinal
  – categories ordered
    • cancer stage I, II, III, IV
    • non-smoker/ex-smoker/smoker
Numeric Data

• Discrete
  – limited # possible values
    • # of children: 0, 1, 2, …

• Continuous
  – range of values
    • weight in lbs: >0
Determining Data Types

- Ordinal (Categorical) vs Discrete (Numeric)
- Ordinal
  - Cancer Stage I, II, III, IV
  - Stage II $\neq$ 2 times Stage I
  - Categories could also be A, B, C, D
- Discrete
  - # of children: 0, 1, 2, …
  - 4 children = 2 times 2 children
Describing Data

• General summary
  – condense information to manageable form
  – numerically
    • Centrality: ‘center’ of data
    • Dispersion: variability
  – graphically/visually

• What analyses are / are not appropriate
Summarizing Categorical Variables

• Display in data table:

<table>
<thead>
<tr>
<th>SubjNo</th>
<th>Smoker</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Never</td>
</tr>
<tr>
<td>2</td>
<td>Current</td>
</tr>
<tr>
<td>3</td>
<td>Never</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>2738</td>
<td>Former</td>
</tr>
<tr>
<td>2739</td>
<td>Former</td>
</tr>
</tbody>
</table>

• Compute #, fraction, or % in each category

Smoking Status:

<table>
<thead>
<tr>
<th></th>
<th>Never</th>
<th>Former</th>
<th>Current</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>621 (23%)</td>
<td>776 (28%)</td>
<td>1342 (49%)</td>
<td>2739</td>
</tr>
</tbody>
</table>
Summarizing Numeric Variables

• Numeric Summaries
  – center of data
    • mean
    • median
  – spread/variability
    • standard deviation
    • quartiles; inner quartile range

• Graphical Summaries
  – histogram, boxplot
Numeric Summaries: Centrality

• Mean
  – average value
  – not robust to outlying values

• Median
  – 50th percentile or quantile
  – the data point “in the middle”
  – ½ observations lower, ½ observations higher
Numeric Summaries: Center

Hospital stays: 3, 5, 7, 8, 8, 9, 10, 12, 35

• Mean
  \[
  \frac{3 + 5 + 7 + \ldots + 35}{9} = 10.78 \text{ days}
  \]
  Only two patients stay > 10.78 days!

• Median
  3  5  7  8  8  9  10  12  35
  Same even if longest stay 100 days! (robust)
Numeric Summaries: Dispersion

• Standard deviation (s, sd)
  – reported with the mean
  – based on average distance of observations from mean
  \[ sd = \sqrt{\text{sum of } (\text{obs} - \text{mean})^2 / (n-1)} \]

Hospital stays: 3, 5, 7, 8, 8, 9, 10, 12, 35
\[ sd = 9.46 \text{ days} \]
  – not robust (sd = 30.9 if longest stay 100 days)
  – sometimes \(~95\%\) of obs w/in 2 sd of mean
  – \emph{at least} 75\%, always!
Numeric Summaries: Spread

• Inner Quartile Range (IQR), Quartiles
  – reported with median (+min, max, quartiles)
  – IQR = distance between 75\textsuperscript{th} & 25\textsuperscript{th} \%iles

\begin{align*}
3 & \quad 5 & \quad 7 & \quad 8 & \quad 8 & \quad 9 & \quad 10 & \quad 12 & \quad 35 \\
\end{align*}

  – 10 − 7 = 3 days
  – range of middle 50\% of data
  – same even if longest stay 100 days! (\textit{robust})
Numeric Summaries

<table>
<thead>
<tr>
<th>Center</th>
<th>Spread</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Std dev (sd)</td>
</tr>
<tr>
<td>Median</td>
<td>IQR</td>
</tr>
</tbody>
</table>

- How to choose which to use? 

  More information?
Graphical Summaries

- **Histogram**
  - divide data into intervals
  - compute # or fraction of observations in each interval
  - good for many observations

Hypothetical cholesterol levels for \( n = 100 \) subjects

![Histogram of Cholesterol Levels](chart.png)
Graphical Summaries: Histogram

- Both means = 2, sd = 1.9, n = 1000
Graphical Summaries: Histogram

- Both means = 2, sd = 1.9, n = 1000
Graphical Summaries: Boxplots

- Graphical extension of median, IQR
- Useful for comparing two variables
- Help identify ‘outliers’
Graphical Summaries: Boxplots

Cholesterol Levels (mg/100mL)
Graphical Summaries: Boxplots

- **Boxplots**
- **IQR**
- **Measurement A**
- **Measurement B**

- Maximum
- “outliers” all greater than Q3 + 1.5 * IQR
- Largest obs less than Q3 + 1.5 * IQR
- Minimum
Data Type & Data Summaries

- Type determines appropriate summary
  - Counts and percents for Categorical
  - Medians for Discrete
  - Means or Medians for Continuous
- Characteristics of data help determine appropriate summary
  - median (IQR) appropriate if skewed, outliers, or small sample
- Summary will help determine appropriate analysis
Missing Data

• A common gap in lectures on describing data regards data that isn’t there.

• Missing data can be categorized as:
  – Missing Completely at Random (MCAR)
    • Chance happening that data is not observed
  – Missing at Random (MAR)
    • Probability of missing data depends on observed variables
  – Non-Ignorably Missing (NMAR)
    • Missing data depends on the unobserved values
Analysis with Missing Data

• Using only the subset of your data that has observed data for the variables of interest is known as a “Complete Case Analysis.”

• At the very least:
  – How much data is missing?
  – Does the probability of missing data depend on other observed variables?
  – Could the probability of missing data depend on the actual values?
Advanced issues in Missing Data

• If missing data is MAR or NMAR, complete case analyses will yield biased results.

• Seek help, but you may hear things like:
  – Available Data analysis (not same as Complete Case)
  – Imputation and Multiple Imputation
    • LVCF or LOCF (Last Value Carried Forward)
    • Propensity Score methods, predictive mean matching, likelihood based methods
  – Weighting Methods
  – Sorry, what you want to estimate is unidentifiable.
Statistical Inference

• Estimation of quantity of interest
  – estimate itself
  – quantify how good an estimate it is

• Hypothesis testing
  – Is there evidence to suggest that the estimate is different from another value?

• Today: proportion, mean, median
Statistical Inference

- Take results obtained in a (random) sample as best estimate of what is true for the population.

- Estimate may differ from the population value by chance, but it should be close.

- How good is the estimate?

- If we took more samples, and got even more estimates, how would they vary?
Statistical Inference

• Ex: proportion of persons in a population who have health insurance, sample size $n = 978$

$$\text{Sample 1} \quad \frac{797}{978} = 0.81$$

We conclude that the estimated % of people with health insurance is 81%.

How variable is our estimate?
Sampling Variability

• Need to know the sampling distribution for the quantity of interest

• One option: take lots of samples, and make a histogram

Sample 2  $\frac{782}{978} = 0.80$
Sample 3  $\frac{812}{978} = 0.83$
Sampling Variability

• Not practical/feasible to keep repeating a study!

• Statistical theory

  – the sampling distributions for means and proportions often look “normal”, bell-shaped

  – from a single sample, can calculate the standard error (variability) of our estimated mean or proportion
Standard Error

- *Standard error (SE)* measures the variability of your *sample statistic* (e.g. a mean or proportion)

- small SE means estimate more precise

- **SE is not the same as SD!**

- SD measures the variability of the sample data; SE measures the variability of the statistic
Standard Error vs Standard Deviation

• Averages less variable than the individual observations

• Averages over larger n less variable than over smaller n

• Ex: heights
  – most people from 5’0” – 6’2”
  – average height for sample of 100 people ~ between 5’5” – 5’7”

• SE ≤ SD!
Inference for Sample Proportions

- Formula for SE of sample proportion

\[ \sqrt{\frac{\text{prop} \cdot (1 - \text{prop})}{n}} \]

- For health insurance example
  - \( n = 978, \ \text{prop} = 0.81 \)
  - \( \text{SE} = \sqrt{(0.81) \cdot (0.19)/978} = 0.01 \)
  - The standard error of the sample proportion is 0.01.
Sampling Distribution of Sample Proportion

- Sample proportions are *normally distributed* if \((\text{prop}) \times (1-\text{prop}) \times n > 10\)
- if \(\leq 10\), then need to use exact binomial (not covered here)

\[
\text{true proportion} - 1.96 \text{ SE} \quad \text{true proportion} \quad \text{true proportion} + 1.96 \text{ SE}
\]

\(~95\%\) of sample proportions within \(~2\) SEs of true proportion
Confidence Interval for Sample Proportion

• 95% Confidence Interval for Sample Proportion

(sample prop – 1.96 * SE, sample prop + 1.96 * SE)

• Interpretation: For about 95/100 samples, this interval will cover the true population proportion

• 95% CI for % population with insurance (n = 978, prop = 0.81)

Check:  
(0.81) * (1-0.81) * 978 = 150.5 > 10

95% CI:  
(0.81 – 1.96 * 0.01, 0.81 + 1.96 * 0.01) =  
(0.79, 0.83)

• For greater confidence (e.g. 99%CI), wider interval
Margin of Error

- In designing a study, you can base your sample size on how wide a confidence interval you want.
- Margin of error reflects how precise you would like to estimate an effect (half the width of the CI).
- Conf. Interval for proportion:
  \[
  \text{Est.} \pm z_{\alpha/2} \text{SE(est.)}
  \]
Margin of Error & Sample Size

N vs C.I. Width by P with C.L.=0.95 C.I. One Proportion

- N=9701
- N=6245
- N=402
- N=264

Z vs p

- 0.500
- 0.800

C.I. Width

10/4/2011

Biostatistics Collaboration Center (BCC)

Lecture 1: Basic Concepts
Standard Error for Sample Mean

• Standard error is the standard deviation of the sample, divided by the square root of the sample size

\[ SE = \frac{SD}{\sqrt{n}} \]

• Ex: \( n = 16 \)
  
  mean SBP = 123.4 mm Hg
  sd = 14.0 mm Hg

• SE of mean = \( \frac{SD}{\sqrt{n}} = \frac{14.0}{\sqrt{16}} = 3.5 \)
Sampling Distribution for a Mean

• Sample means follow a t-distribution IF
  – underlying data is (approximately) normal
  \textit{OR}
  – \( n \) is large

• A sample mean from a sample of size \( n \) with
  have a t distribution with \( n-1 \) degrees of freedom

• \( t_{n-1} \)
For a sample of size $n = 16$, the sample mean would have a $t_{15}$ distribution, centered at the true population mean and with estimated standard error $sd/\sqrt{n}$. 
Confidence Interval for Sample Mean

- Ex: \( n = 16 \)
  \[ \text{mean SBP} = 123.4 \text{ mm Hg} \]
  \[ \text{sd} = 14.0 \text{ mm Hg} \]

- SE of mean = \( \frac{SD}{\sqrt{n}} = \frac{14.0}{\sqrt{16}} = 3.5 \)

- Assume that SBP is approximately normally distributed (at least symmetric, bell-shaped, not skewed) in the population

- 95% CI for sample mean
  \[ \text{mean} \pm 2.131 \times \text{SE} = 123.4 \pm 2.131 \times 3.5 \]
  \[ = (115.9, 130.8) \text{ mm Hg} \]
Confidence Interval for Sample Mean

- Suppose instead mean and sd were for sample of size $n = 100$

- 95% CI for sample mean
  - $n = 100$
    \[
    \text{mean} \pm 1.984 \times \frac{\text{sd}}{\sqrt{n}} = 123.4 \pm 1.984 \times \frac{14}{\sqrt{100}} = (120.6, 126.2) \text{ mm Hg}
    \]
  - $n = 16$
    \[
    \text{mean} \pm 2.131 \times \frac{\text{sd}}{\sqrt{n}} = 123.4 \pm 2.131 \times \frac{14}{\sqrt{16}} = (116.5, 130.3) \text{ mm Hg}
    \]
Confidence Interval for Sample Mean

- Suppose instead mean and s were for sample of size n = 100

95 % CI for sample mean

- $n = 100$
  \[ \text{mean} \pm 1.984 \times \frac{s}{\sqrt{n}} = 123.4 \pm 1.984 \times \frac{14}{\sqrt{100}} \]
  \[ = (120.6, 126.2) \text{ mm Hg} \]

- $n = 16$
  \[ \text{mean} \pm 2.131 \times \frac{s}{\sqrt{n}} = 123.4 \pm 2.131 \times \frac{14}{\sqrt{16}} \]
  \[ = (116.5, 130.3) \text{ mm Hg} \]
Confidence Interval for a Sample Mean

• Confession: it’s never incorrect to use a t-distribution (as long as underlying population normal or n large)

BUT

as n gets large the t-distribution and normal distribution become indistinguishable
Margin of Error and Means

N vs Dist. to Limit by S with C.L.=0.95 C.I. One Mean

- N=984
- N=753
- N=112
- N=87

Z vs Dist. to Limit:
- 14.000
- 16.000
Hypothesis Testing

• Confidence intervals tell you
  – best estimate
  – variability of best estimate

• Hypothesis testing (i.e. proof by contradiction)
  – Is there really a difference between my observed value and another value?
Hypothesis Testing

• From our sample of $n = 978$, we estimated that 81% of the people in our population had health insurance.

• From previous census records, you know that 10 years ago, 78.5% of your population had health insurance.

• Has the percent of people with insurance really changed?
Hypothesis Testing

• Suppose the true percent with health insurance is 78.5%
  – Called the *null hypothesis*, or $H_0$

• What is the probability of observing, for a sample of $n = 978$, a result as or more extreme than 81%, given the true percent is 78.5%?
  – Called the *p-value*, computed using normal distribution for sample proportions (and t-distribution for sample means)

• If probability is small, conclude that supposition might not be right.
  – Reject the null hypothesis in favor of the alternative hypothesis, $H_a$
    (opposite of the null hypothesis, related to what you think may be true)

• If the probability is not small, conclude do not have evidence to reject the null hypothesis
  – Not the same as ‘accepting’ the null hypothesis, or showing that the null hypothesis is true
Hypothesis Testing

H₀: True proportion is 78.5%
Hₐ: True proportion is not 78.5%

Supposed true proportion = 0.785
P-value = Prob(sample prop > 0.81 or sample prop < 0.76 given true prop = 0.785) = 0.012
Hypothesis Testing

• It is not very likely ($p = 0.012$) to observe our data if the true proportion is 78.5%.

• We conclude that we have sufficient evidence that the proportion insured is no longer 78.5%.

• “Two-sided” test: observations as extreme as $0.81 > 0.81$ or $< 0.76$

• “One-sided” test: observations larger than 0.81
  – Has the percent of people with insurance really increased?
  – $H_a$: True proportion greater than 78.5%
  – $p$-value = 0.006 (only area above 0.81)
  – Only ok if previous research suggests that the prop is larger
Hypothesis Testing

- Similar for sample means, except that p-values are computed using t-distribution, if appropriate.

- \( n = 16 \) with \( \text{SBP} = 123.4 \), \( \text{sd} = 14.0 \) mm Hg

- Previous research suggests that population may have lifestyle habits protective for high BP.

- Do we have evidence to suggest that our population has SBP lower than 125 mm Hg (prehypertensive)?
Hypothesis Testing

H₀: True mean is 125 mm Hg
Hₐ: True mean is less than 125 mm Hg

• Using \( t_{15} \) distribution, we find a one-sided \( p \)-value of 0.32

• This too large to reject the null hypothesis in favor of the alternative.
  – Usually need \( p < 0.05 \) to “reject null” in favor of alternative
Misinterpretations of the p-value

- A p-value of 0.32 (or > 0.05) DOES NOT mean that we “accept the null”, or that there is a 32% chance that the null is true
  - we haven’t shown that the SBP of this group is equal to 125
  - true value may be 124.5 or 125.1 mmHg

- Can only “reject the null in favor of the alternative” or “fail to reject the null”

- If you “fail to reject” that doesn’t mean the alternative isn’t true. You may not have had a large enough n!
Distribution-Free Alternatives

- Recall for the normal distribution to be useful for sample proportions, we need $n \times \text{prop} \times (1-\text{prop}) > 10$
  - If not, use exact binomial methods (not covered here)

- For the t-distribution to be useful for sample means, we need the underlying population to be normal or $n$ large
  - What if this isn’t the case?
    - data are skewed or $n$ is small
    - e.g. length of hospital stay
  - Use nonparametric methods
    - look at ranks, not mean
    - the median is a nonparametric estimate
Nonparametric Methods

- nonparametric methods are methods that don’t require assuming a particular distribution

- a.k.a distribution-free or rank methods

- nonparametric tests well suited to hypothesis testing

- not as useful for point estimates or CIs

- especially useful when data are ranks or scores
  - Apgar scores
  - Vision (20/20, 20/40, etc.)
Nonparametric Methods for Sample Median

• Alternative to using sample mean & t-distribution
• Instead do inference for median value
  – CI for median uses ranks
    • $n = 16$, ~ 95% CI for median
      (4\text{th} largest value, 13\text{th} largest value)
    • $n = 25$, ~ 95% CI for median
      (8\text{th} largest value, 18\text{th} largest value)
    • necessary ranks depend on $n$ and confidence level
      Stat packages or tables
  – Hypothesis test for one sample: Sign Test or Wilcoxon Signed Rank test
Wilcoxon Signed Rank Test

• A roughly normally distributed sample (n=16) of SBP (mean=123.4, sd=14) was used to test if the median=125. p-value (provided by statistics package) = 0.570

• Skewed data, though was also created with same mean and sd, p=0.006

• Typically when parametric assumptions of the data are met, non-parametric tests are 95% as powerful as their corresponding parametric test. When assumptions are not met, np tests can be more powerful!
Parametric vs. Nonparametric

• Nonparametric always okay
• Nonparametric more conservative than parametric
  – e.g. 95% CIs for medians sometimes twice as wide as those for the mean
• If your data satisfy the requirements, or n is fairly large, probably best to use parametric procedures
Useful References

• Intuitive Biostatistics, Harvey Motulsky, Oxford University Press, 1995
  – Highly readable, minimally technical

• Fundamentals of Biostatistics, Bernard Rosner, Duxbury, 2000
  – Useful reference, somewhat technical

  – Useful reference for analysis write-ups and journal reviews
Next Time

Two group comparisons
Contact BCC

• Please go to our web site at:

http://www.feinberg.northwestern.edu/depts/bcc/

• Fill out our online request form for further collaboration!