Biost 517 / Biost 514 Applied Biostatistics I / Biostatistics I

Timothy A. Thornton, Ph.D.
Associate Professor of Biostatistics
University of Washington

Lecture 3:

Means and medians cont.; Measuring spread;
Example with univariate descriptive statistics

Harmonic Mean

- The harmonic mean is related to the mean of the reciprocal of the data

$$\frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{X_i}}$$

- Harmonic mean has scientific interpretation in some applications
  - E.g., in electricity, resistance of parallel resistors
  - E.g., in studying vascular flow and blood pressure

Descriptive and Statistical Uses of: Mean and Median

(Arithmetic) Mean: Numerical Variables

- Defined only for variables that take on numeric values (sum must make sense)
- Most sensible when differences have scientific interpretation on a constant scale
  - (5 - 4) should be equivalent to (3 - 2), etc.
  - (But see comments regarding comparisons of ordered categorical variables)

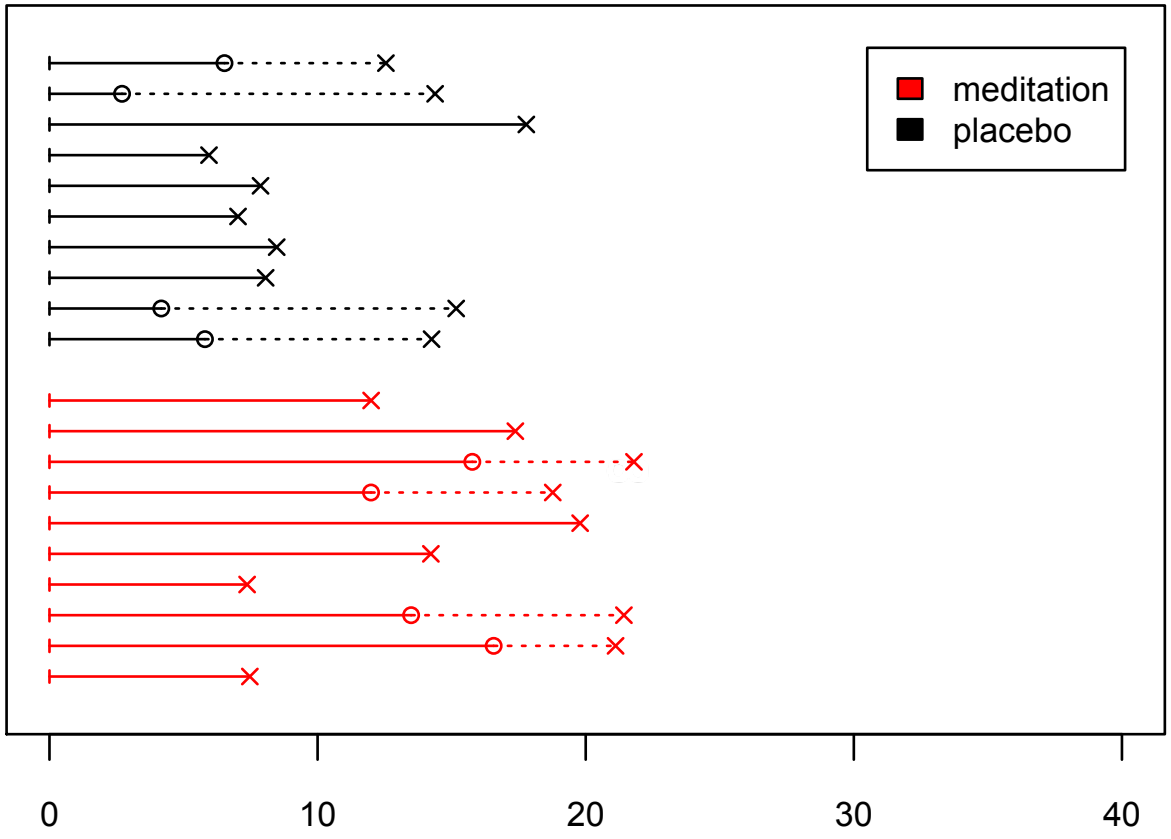
Mean: Ordered Categorical Data

- The mean by itself is not scientifically interpretable, which is why we do not use it descriptively
- The mean can still detect differences in the distributions
  - Sensitive to certain tendencies for higher measurements in one group

Right Censored Data

- It is common to not have been able to exactly observe the time to event for all study participants.
- Why so? Most common reasons are:
  - the study ended (say, after 30 weeks) and some participants had not yet had an event (e.g., relapse); (administrative censoring)
  - the participant left the study before having a relapse; (loss to follow-up)
- These all lead to **right-censored data**.
- A special type of missing data: the exact value is not always known
  - Some measurements are known exactly
  - Some measurements are only known to exceed some specified value (perhaps different for each subject)

Right Censored Data



- Right-censored data: on a graph, the unobserved event time would lie somewhere to the **right** of the censoring time

Right Censored Data

Unobserved :

True times to event : $\left\{T_1^0, T_2^0, \ldots, T_n^0\right\}$

Censoring Times : $\left\{C_1, C_2, \ldots, \mathrm{C}_n\right\}$

Observed data :

Observation Times : $T_i = \min\left(T_i^0, C_i\right)$

Event indicators : $\mathrm{D}_i = \begin{cases} 1 & \text{if } T_i = T_i^0 \\ 0 & \text{otherwise} \end{cases}$

Mean for censored data?

- For a censored variable, the observation time is a mixture of times to event and times to censoring
- The mean for a censored variable is not scientifically scientifically interpretable, which is why we do not use it descriptively
- We must know all values in order to compute the mean
- We will later introduce special (Kaplan-Meier) methods to describe censored variables

Mean: Descriptive Uses

- Characterizing sample
  - Often used as a “typical value”
  - As previously mentioned, mean is heavily influenced by extreme values
    - Descriptively, this might be undesirable. Suppose a disease kills most people quickly but a small fraction recover completely. Then “mean survival” is not a useful summary to share with patients
    - Sometimes this is desirable. A popular restaurant considering an expansion will be interested in “mean profit per table” even if there are extreme values.

Mean: Statistical Uses

- Assessing validity of assumptions
  - Often we study the association between a predictor of interest and a response by examining whether the mean response varies with the predictor
  - Best measure of association to summarize potential confounding

Median: Type of Variables

- Concept of a median is defined for any ordered variable
- Special methods (Kaplan-Meier estimates) must be used with censored data
  - The simple sample median is not of interest
    - The observation time is a mixture of times to event and times to censoring
  - More often able to estimate median from censored data (using Kaplan-Meier methods) than mean

Median: Descriptive Uses

- Characterizing distribution of sample
  - "Typical" value, especially when we don't want the measure of central tendency to be sensitive to a few unusually extreme observations.
- Because the median is less responsive to extreme values than the mean, the difference between the median and mean characterizes the skewness of a distribution

Ordering of Means, Median

- For positive random variables:
  - Arithmetic mean > Geometric mean > Harmonic mean
  - Median can be anywhere in this ordering

Univariate Measures of Spread

Descriptive Statistics: Measuring the Spread of a Distribution

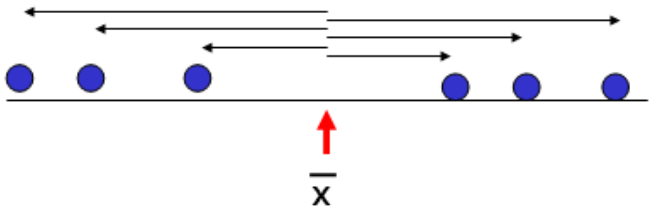
- Commonly used descriptive statistics that measure dispersion, or how variable the data include:

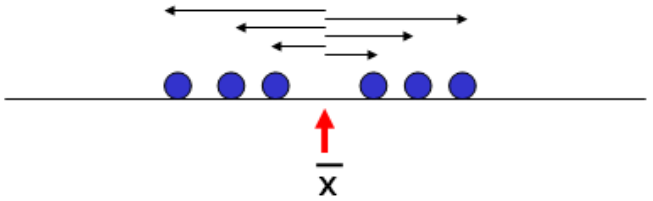
1. Variance & Standard deviation
2. Inter-quartile range
3. Range

Sample Variance

- The **sample variance**, s^2, is the arithmetic mean of the squared deviations from the **sample mean**:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$





Sample Standard Deviation

- The **sample standard deviation (SD), s,** is the square-root of the sample variance

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

- **s is in** the same units as the original random variable **x**

Why divide by n-1 for the sample variance

- Division by n-1 instead of n in the sample variance calculation is a common cause of confusion.

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

- Why n-1? Note that $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$
- Thus, if you know any n-1 of the differences, and you also know $\bar{x}$, then you can figure out the nth difference.
- Thus the number of "freely varying" observations, n-1 in this case, is called the "degrees of freedom."
- Sample variance is an unbiased estimator of the population variance

Variance: Interpretation

- Squaring accentuates large errors
  - More convenient mathematically than absolute value
  - Variance measured in units^2
- Variance has theoretical basis as second central moment of a distribution. More on this later.
- The variance is important for sampling distributions used in statistical inference
  - Variance is a fundamental parameter of the Normal distribution

Variance: Types of Variables

- Appropriate for numeric variables having interpretable differences
- Do not use with censored variables

Variance: Used in Practice

- Larger sample variance means more variable measurements. However:
- Variance is not very useful as a descriptive statistic because units are the square of the units of the variable
  - Variance of height in *cm* has units cm^2

- Assessing validity of models
  - Many classical analysis methods rely on assumptions about within-group variances

Variance: Scientific Questions

- Quantifying or comparing distributions
  - Sometimes we are scientifically interested in the spread of the distribution
    - E.g., Does treatment A for diabetes stabilize blood sugar better than treatment B?
  - Later in this course we will see that the sampling distribution of the variance in large samples is known
    - But it takes larger sample sizes than is required for inference about the mean

Example: Creatine kinase (CK) activity

- We previously calculated the mean of **creatine kinase** (**CK**) activity to be

$$\bar{x} = \frac{121 + 82 + 100 \cdots + 95 + 42}{36} = \frac{3538}{36} = 98.277$$

121	82	100	151	68	58
95	145	64	201	101	163
84	57	139	60	78	94
119	104	110	113	118	203
62	83	67	93	92	110
25	123	70	48	95	42

Sample Variance and SD of CK activity

- The sample variance and sample standard deviation of **CK** activity is

$$s^2 = \frac{(121-98.28)^2 + (82-98.28)^2 + (121-98.28)^2 \cdots (42-98.28)^2}{36-1} = \frac{57071.22}{35} = 1630.606$$

$$s = \sqrt{1630.606} = 40.381$$

Mean Deviation

- Definition
  - The average absolute distance from the mean

$$\frac{1}{n}\sum_{i=1}^{n}\left|X_i - \overline{X}\right|$$

- Defined whenever the SD/variance is defined
- Possible alternative to SD /variance, however:
  - Absolute values are harder to work mathematically
  - Sampling distribution is thus harder to derive
  - Not often used as it not very useful for statistical inference

Range

- Two definitions:
  - (Minimum, maximum) values
  - Maximum - minimum is used by some people

Range: Types of Variables

- Only makes sense for ordered variables

- Not appropriate for (right) censored time-to-event
  - If the earliest event happens before any censoring, then that is the minimum, but if the earliest event is a censoring event, then we don't know the minimum
    - If the earliest event is a censoring event, then minimum is (interval) censored
  - The maximum event time is only known among uncensored observations
  - Don't use min/max/range; Instead use Kaplan-Meier curves to estimate survival curve (Lecture 8)

Range: Used in Practice

- In describing a sample, shows the breadth of data acquired
  - Might detect errors in data collection or data entry
    - Values out of range
- We don't use the min/max/range to summarize distributions or compare distributions across groups
  - The sample min/max/range depend heavily on the sample size

Range: Types of Variables

- Only makes sense for ordered variables
- Not appropriate for (right) censored time-to-event
  - If the earliest event happens before any censoring, then that is the minimum, but if the earliest event is a censoring event, then we don't know the minimum
    - If the earliest event is a censoring event, then minimum is (interval) censored
  - The maximum event time is only known among uncensored observations
  - Don't use min/max/range; Instead use Kaplan-Meier curves to estimate survival curve (which will be discussed later in the course)

Range: Used in Practice

- In describing a sample, shows the breadth of data acquired
  - Might detect errors in data collection or data entry
    - Values out of range
- We don't use the min/max/range to summarize distributions or compare distributions across groups
  - The sample min/max/range depend heavily on the sample size

Interquartile Range (IQR)

- We previously introduced the interquartile range (IQR)
- Two definitions:
  - 25^{th} (Q1), 75th (Q3) percentiles of sample
  - Difference between quartiles is also used
    - IQR=Q3 - Q1, and covers the values falling within the middle 2 quartiles (the center 50%) of the distribution.
- IQR is also used to indicate the variability around the **sample median, Q2**
- Only makes sense for ordered variables
- Do not use with censored variables

IQR: Used in Practice

- Characterizing the sample
  - A measure of spread that is insensitive to outliers
  - Central 50% of the data
- Assess outliers in the data, i.e., 1.5 x IQR rule
- Rarely useful for assessing the validity of statistical assumptions. While some statistical methods assume equal spread of distributions, these assumptions tend to be about variances.

Example: Use of Univariate Descriptive Statistics

Standard Univariate Description

- Often easier to ask for standard descriptive statistics on all variables….
  - Sample size
  - Number of missing
  - Mean
  - Standard Deviation
  - Minimum
  - 25th percentile
  - Median (50th percentile)
  - 75th percentile
  - Maximum
- … and then consider which statistics are meaningful and relevant

Example: Prostate cancer (PSA dataset)

- Prognostic value of nadir PSA on time in remission from prostate cancer
- PSA data set: 50 men who received hormonal treatment for advanced prostate cancer
- Followed at least 24 months for clinical progression, but exact time of follow-up varies
- Nadir PSA: lowest level of serum prostate specific antigen achieved post-treatment

Example: PSA Data Variables

- Prognostic value of PSA in hormonally treated prostate cancer
  - ptid Patient ID
  - nadir Lowest PSA following treatment
  - pretx Pre-treatment PSA
  - ps Performance status (0 – 100)
  - bss Bone scan score (1, 2, or 3)
  - grade Tumor grade (1, 2, or 3)
  - age Age (years)
  - obstime Time until relapse or end of study (months)
  - inrem In remission at obstime

Ex: PSA Descriptive Statistics

	n	ms	mean	stdev	min	25%le	mdn	75%le	max
ptid	50	0	25.5	14.6	1.0	13.2	25.5	37.8	50
nadir	50	0	16.4	39.2	0.1	0.2	1.0	9.5	183
pretx	43	7	670.8	1287.6	4.8	52.0	127.0	408.0	4797
ps	48	2	80.8	11.1	50.0	80.0	80.0	90.0	100
bss	48	2	2.5	0.7	1.0	2.0	3.0	3.0	3
grade	41	9	2.2	0.8	1.0	2.0	2.0	3.0	3
age	50	0	67.4	5.8	58.0	63.2	66.0	70.0	86
obstime	50	0	28.5	18.4	1.0	12.5	28.0	42.0	75
inrem	50	0	0.3	0.4	0.0	0.0	0.0	1.0	1

Example: PSA Data Variables

- Types of data
  - ptid Unordered categorical (coded as numbers)
  - nadir Continuous (ratio)
  - pretx Continuous (ratio)
  - ps Continuous (ratio) (measured discretely)
  - bss Ordered categorical
  - grade Ordered categorical
  - age Continuous (ratio)
  - obstime Censored continuous
  - inrem Binary indicator of censoring for obstime

Example: PSA Data Variables

- Relevant univariate statistics
  - ptid none
  - nadir Mean, SD, Min, Max, Quantiles
  - pretx Mean, SD, Min, Max, Quantiles
  - ps Mean, SD, Min, Max, Quantiles
  - bss Min, Max, Quantiles (Frequencies)
  - grade Min, Max, Quantiles (Frequencies)
  - age Mean, SD, Min, Max, Quantiles
  - obstime (Kaplan-Meier estimates needed)
  - inrem (Kaplan-Meier estimates needed, though mean of inrem tells us proportion of uncensored observations)

Example: Relevant Univariate Statistics

	Obs	ms	mean	stdev	min	25%le	mdn	75%le	max
ptid	50	0							
nadir	50	0	16.4	39.2	0.1	0.2	1.0	9.5	183
pretx	43	7	670.8	1287.6	4.8	52.0	127.0	408.0	4797
ps	48	2	80.8	11.1	50.0	80.0	80.0	90.0	100
bss	48	2			1.0	2.0	3.0	3.0	3
grade	41	9			1.0	2.0	2.0	3.0	3
age	50	0	67.4	5.8	58.0	63.2	66.0	70.0	86
obstime	50	0							
inrem	50	0							