

STAT 301 Midterm Review

First Draft

by Jingci Meng

last update: July 23, 2006

1 Some Basic Concepts

- A statistics **Population** is the set of measurements (or record of some qualitative trait) corresponding to the **entire collection** of units about which information is sought.
- A **Sample** from a statistical population is the **subset** of measurements that are actually collected in the course of an investigation.

2 Data, Graphical Displays, Numerical Summizations

- Two data types:
 - (1) Categorical data (unordered or ordinal)
 - (2) Numerical data (discrete or continuous)
- Bar/Pareto chart, Pie chart, Stem and leaf plot, dotplot, boxplot
- Dot plot
Horizontal line represents the value of the observations. Vertical line represents the number of counts.
- Histogram
There are three versions of the histogram :
 - (i) Frequency histogram : Height = Frequency
 - (ii) Relative frequency histogram : Height = Relative frequency
 - (iii) Density scale histogram : Height = Relative frequency / Width

Note :

 - (i) For a density scale histogram, the area = the relative frequency.
 - (ii) If all the class intervals do not have the same width, use a density scale histogram.
 - (iii) The left tail is longer than the right, the distribution is called skew to the left. If the right tail is longer than the left, the distribution is called skew to the right.
- Stem and leaf plots
- Outlier
 - Outlier is any value that is markedly different from the main body of values in the distribution.
 - Dot plot is a good way to check for outliers.(sometimes subjective)

- Measure of center

Sample mean: $\bar{x} = \frac{x_1+x_2+\dots+x_n}{n}$

Sample median: Sort the data first

- if number of observation is even, then it is the average of the middle two numbers.
- if number of observation is odd, then it is the middle one. Note :
 - (i) If a distribution is approximately symmetric, the values of sample mean and median will be similar.
 - (ii) Sample mean $>$ sample median for a skewed to the RIGHT distribution.
Sample mean $<$ sample median for a skewed to the LEFT distribution.
- Measure of spread
Sample range: denoted by R, the largest value minus the smallest

Interquartile range: denoted by IQR.

IQR = Q3-Q1

Note Q1=25th Percentile, Q3=75th Percentile. where 100p-th Percentile=

- * if np is not an integer, round it up to the next integer and find the corresponding ordered value;
- * if np is an integer, calculate the average of the np-th and (np+1)th ordered values.
- Standard deviation of sample
Deviation of one observation $x-\bar{x}$ (sum of deviation equal 0)
Sample variance $s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$
Sample standard deviation $s = \sqrt{s^2}$
- Box plots
How to identify first quartile,third quartile,median,smallest point,largest point,range.

3 Probability Theory

- **Sample space** S : the collection of all possible outcomes.
- **Events** : subsets of sample space
- **Equally likely case:** If there are k elementary outcomes in S , and an event A consisting of m elementary outcomes, then

$$P(A) = \frac{m}{k}$$

- Useful laws:
 1. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
 2. $P(A) = 1 - P(A^c)$

3. $P(A) = P(AB) + P(A\bar{B})$
- If A and B are **mutually exclusive**, $P(A \cap B) = 0$.
 - $(A \cup B)^c = A^c \cap B^c$ (i.e. $\bar{A}\bar{B}$)
 - $P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$
 - $P(A) = P(AB) + P(A\bar{B})$
 - Conditional Probability: $P(A|B) = P(AB)/P(B)$, or equivalently, $P(AB) = P(B)P(A|B)$.
 - Independence: Two events A and B are independent if $P(A|B) = P(A)$. Equivalent conditions are $P(B|A) = P(B)$ or $P(AB) = P(A)P(B)$.