



Basic Biostatistics

Session One



Terms to Know

Greek Letter	Name	Statistical Term
α	Alpha	Type I error
β	Beta	Type II error
δ	Delta	Difference
μ	Mu	Population Mean
σ	Sigma	Population Standard Deviation
\bar{x}	x bar	Sample mean
s	s	Sample standard deviation
n	n	Sample size



Doc I'm not feeling well?

- ◆ What is normal/ healthy?
- ◆ What is sick?
- ◆ How do you tell them apart in a systematic way?
 - Not just because Mom said you're sick
 - Not always black and white
- ◆ Statistics first role is to differentiate well from unwell.
 - Studies about Harm, Diagnosis, Screening all depend on the ability to differentiate the two groups from each other.
 - Answering the Clinical Question begins here.



Sample versus Population

- ◆ The population defines normal.
 - Did anybody see the Twilight Zone episode where everybody has pig noses and the “beautiful girl” is ugly??
 - Normal is defined by its surroundings
- ◆ Best approach is to compare your patient to all similar individuals in the population.
 - This is practically impossible
 - Too many
 - Too expensive
 - Too much work!
 - Create a Sample population to compare

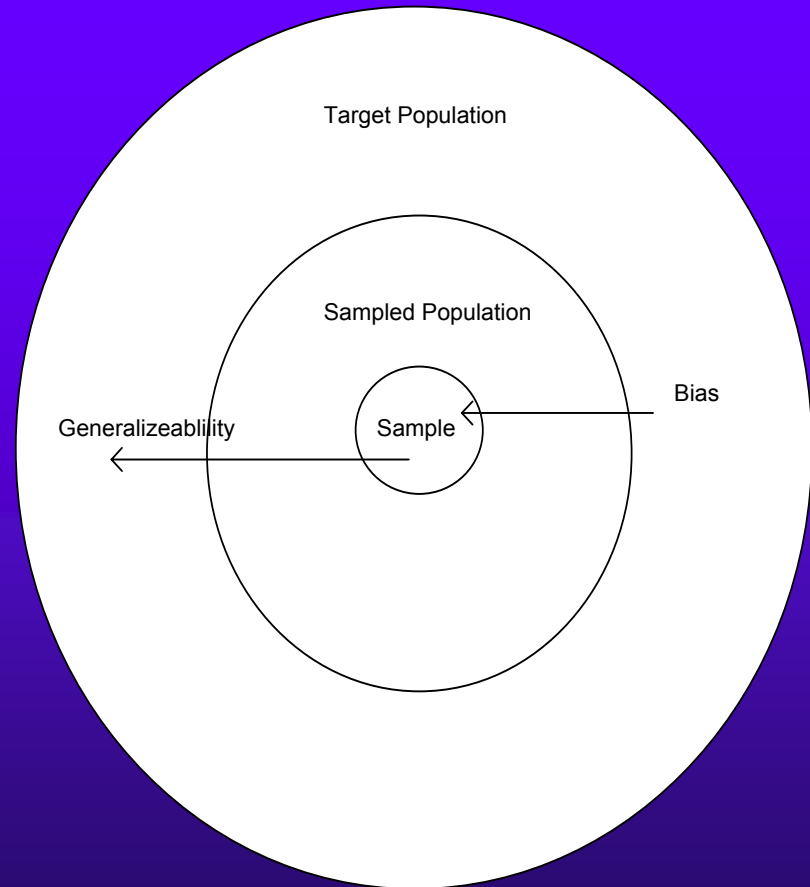


Sample

- ◆ Create a smaller version of the population to compare with your patient.
- ◆ Be sure it truly represents the population
 - Best approach is a big sample drawn at random
 - Usual approach is a small group and hope it is close to representative.

Not all samples are created equal.

- ◆ Statistics tell us about the relationships between these groups.
- ◆ They do not correct for systematic bias.
 - If the study design is flawed then...
 - Does anyone remember the post election polls?



A vertical image on the left side of the slide shows a metal key with a circular head and a notched bit, resting on a light-colored, textured surface like sand or gravel.

Back to terms

Variable: the event or characteristic under study.

- **Dependent:** the outcome you are interested in
 - The normal range for hemoglobin
 - Hemoglobin on the X axis
 - The range of **cholesterol** seen in army recruits
- **Independent:** the agent/event you believe might effect the dependent variable. There can be more than one.
 - Does smoking effect the serum cholesterol
 - Cholesterol on the X axis, **Smoking in Packs/Day on the Y**
 - Does salt intake effect systolic blood pressure
 - Systolic BP on the X, **Na intake in mg on the Y**



Types of variables

- ◆ The type of variable you wish to study
 - Determines the Statistical Test used to differentiate normal from abnormal
 - Defines the study type needed to answer the clinical question
- ◆ Frames the question
 - Yes/No
 - Honors/High Pass/Pass/Fail
 - 0,1,2,3,4... 100



Types of variables

◆ Discrete

– Nominal

- Named categories with no order
 - Gender, blood type, marital status, live/dead, healthy/sick

– Ordinal

- Ordered Levels, but differences are not measurable
 - Stage of cancer, performance status, class of heart failure


◆ Continuous

– Interval

- Equal distance between points but no real zero point
 - Quality of life score, IQ

– Ratio

- Equal distance between points with a real zero
 - Cholesterol, Hemoglobin, age, weight



Still can't answer what is normal...yet.

- ◆ A sample is used to estimate what one would expect in the vast majority of the population
 - You can't measure the entire country's hemoglobin to reach the normal curve
 - Estimate with census data
- ◆ Even more
 - Biologic variables vary among and within individuals: there is not one normal hemoglobin value
 - So is normal is fuzzy, what does that make abnormal?



Statistics to the Rescue!

- ◆ Central Tendency can help.
 - The fact that variables focus around an average value and we can define this and describe the pattern of values around the average
 - Mean, Median, Mode, Variance, Deviation...



M and M and M

- ◆ Mode: the most frequently occurring value in a sample
 - There can be one or more than one mode
- ◆ Median: the value in the middle of ranked data, the 50th percentile
 - Half the values above and half below (split the difference if even number in the sample)
- ◆ Mean: the average, \bar{x} (the sum of all values divided by the number in the sample)



For Example:

Blood Glucose data from healthy “volunteers” in the medical school with ages ranging from 22 to 30.

78,82,82,98,100,130

Mode=

Median=

Mean=



For example:

Blood Glucose data from healthy “volunteers” in the medical School with ages ranging from 22 to 30.

78,82,82,98,100,130

Mode= 82

Median= 90 (split difference between 82 and 98)

Mean= sum of all values above / 6 = 95

Problem is the sample is derived from only 6 people



Measures of Dispersion

- ◆ How does the data occur around the mean, median and mode.
 - Min-max: lowest and highest values
 - Range: difference between the highest and lowest values
 - Variance: the average value of the squared deviations of each value from the mean; (s^2)
 - Standard Deviation: the square root of the variance; (s)



For example:

Blood Glucose data from healthy “volunteers” in the medical School. Ages range from 22 to 30.

78,82,82,98,100,130

Mode= 82

Min-max:

Median= 90

Range:

Mean= 95

Variance:

Standard Deviation:



For example:

Blood Glucose data from healthy “volunteers” in the medical school Ages range from 22 to 30.

78,82,82,98,100,130

Mode= 82

Min-max: 78,130

Median= 90

Range: 52

Mean= 95

Variance: $[(17)^2 + (13)^2 + (-3)^2 + (-5)^2 + (-36)^2] / 6 = 877$

Standard Deviation: $\sqrt{877} = 30$



Dispersion

- ◆ How much variability is there in the population and how our sample relates to the “real” value.
 - ◆ These don’t include measures of:
 - Precision
 - Accuracy
-
- I never forget the bull’s eye example.
 - Precision around and single point.
 - Accuracy around the center.
 - They don’t have to occur together
 - Three bad shots all in one spot is very precise but not accurate.
 - Three shots all very close but spread around the center is accurate but not precise.

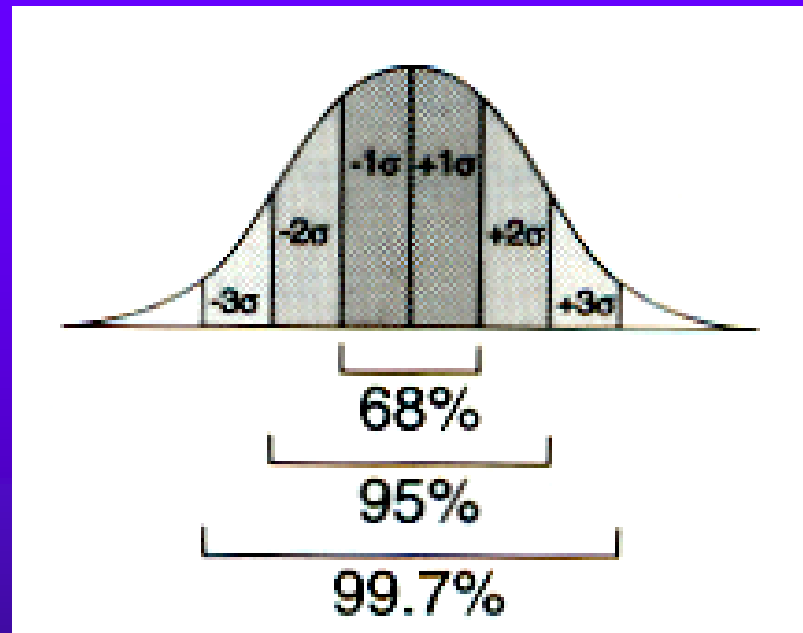


Again to our question?

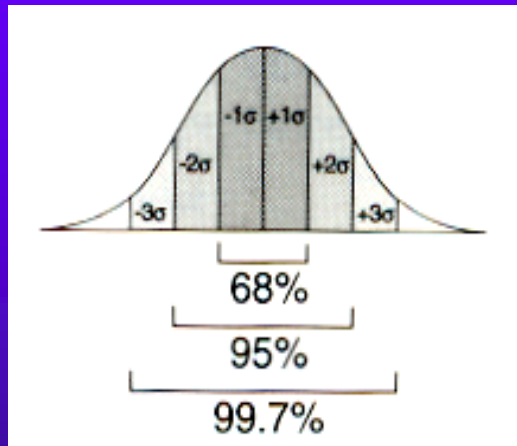
What is normal and what is sick?

- Biologic variables usually follow a normal distribution.
- Normal (Gaussian) distributions have statistical characteristics that make frequencies (probability) of a value predictable.

Gaussian or Normal Distribution Curve

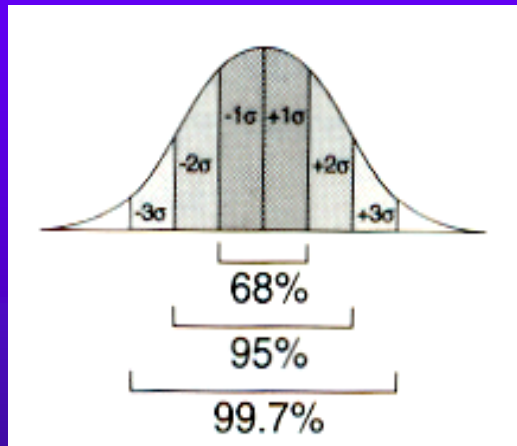


Characteristics of the Normal Curve



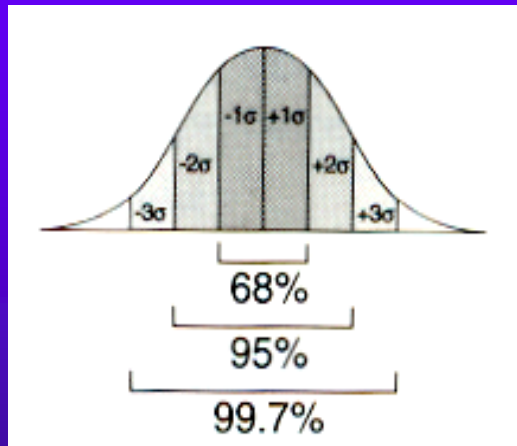
- ◆ **Unimodal**
 - One mode only
- ◆ **Symmetric**
 - 50% of the values occur above and below the mean and median
- ◆ **Asymptotic**
 - Never reaches 0 or 1
 - The curve is a probability not a value. (not real numbers)
- ◆ **Mean, Median and Mode are the same**

Characteristics of the Normal Curve



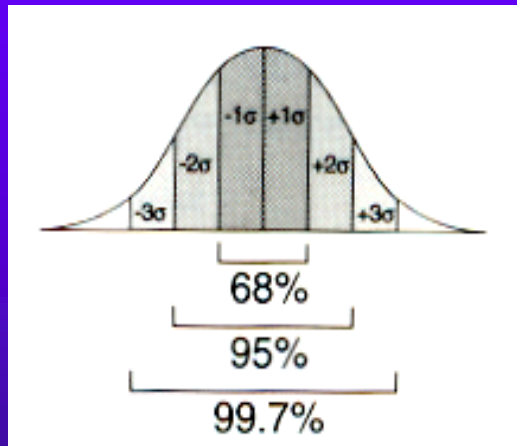
- ◆ Values on the horizontal axis are Z values ranging from $0 <$ to < 1 (probability units)
- ◆ The mean is the center and the values in Standard Deviations account for proportions of the population
 - 1 SD = 68% of the sample
 - 2 SD = 95% of the sample
 - 3 SD = 99% of the sample

Characteristics of the Normal Curve



- ◆ The SD or σ is the amount of the population encompassed by those values
- ◆ 1 SD around the mean is 68% of individuals have a value inside while 32% lie outside
 - Half of the 68% is above the mean/median, half below.
 - You can define normal in this curve.

Characteristics of the Normal Curve



- ◆ This curve can represent variation in a sample
 - Standard deviation
- OR
- ◆ Variation of samples around the population
 - Standard Error
- ◆ A single group compared within itself
 - Standard Deviation
- ◆ Multiple Groups compared to each other
 - Standard Error



Gaussian Distribution

- ◆ Many biologic variables follow this pattern
 - Hemoglobin, Cholesterol, Serum Electrolytes, Blood pressures, age, weight, height
- ◆ One can use this information to define what is normal and what is extreme
- ◆ In clinical medicine 95% or 2 Standard deviations around the mean is normal
 - Clinically, 5% of “normal” individuals are labeled as extreme/abnormal
 - We just accept this and move on.



Back to our patient

- ◆ His serum glucose is 124?
 - Is this healthy or abnormal
 - Use the characteristics of the normal distribution to tell us what the chance is that this value or one more extreme occurs in the population
 - Simply, what percent of the sample has a value equal to or greater than this value.



So the Glucose is 124

- ◆ If the Mean glucose in your sample is 100
- ◆ And if ...34% of individuals have glucose readings from 100-110 and 47% individuals have readings from 100 to 120.
 - Standard deviation was 10
- ◆ What's Normal?

So does 124 fall within the defined criteria for a normal blood glucose.



Answer

- ◆ Mean 100
- ◆ Standard deviation is
 - If 34% have sugars between 100-110 then 68% around the mean 90-110 is 1 SD or 10.
 - Two standard deviations is 20:
 - Therefore, normal blood glucose is 80 to 120.
 - Two standard deviations above and below the mean.



Answer

◆ So....

- Practically, two standard deviations encompass 95% of the population (normal blood glucose levels) and $\frac{1}{2}$ of that amount or 47.5% occur above and 47.5% occur below.
- So your mean plus 2 standard deviations will give you the upper limit of normal and your mean minus 2 standard deviations will give you the lower end of normal



Normal Values

- ◆ So we know that 5% of blood glucose of normals are readings greater to or less than 80-120 (we accept this fact and move on)
- ◆ But where are the values
 - 2.5% are above and 2.5% below the range
 - The outlining values are extreme or abnormal/unwell/unaffected by therapy or risk, etc...
 - A P value can be used statistically to say how confident you are in differentiating the two groups.



P values

- ◆ The probability of a specific value or a more extreme value in the given population
- ◆ In clinical medicine a p value of less than or equal to 0.05 (5%) is usually considered statistically significant
 - A P value states chance that the relationship between our variables is a “fluke” or
 - $p=0.05$, a 5% chance that the values compared are not different statistically (overlap between two distributions)
 - 5% of all values are further from the central point on the Distribution curve.(normal ranges for a single value)



If you care.....

We can tell for an individual value in sample where it falls on the Gaussian curve

- Z score will tell

$$Z = \frac{\bar{X} - (X / SD)}{\bar{X} - (\mu / \sigma)}$$

For this example: $124 - 100 / 10 = 2.4$

- Z score for 2.4 shows in a published table that 1% of the values in the population are more extreme
- P value for this glucose is $p=0.01$
- 1% chance that this glucose is not abnormal