BE.104 Spring
Biostatistics: Distribution and the Mean
J. L. Sherley

Outline: 1) Review of Variation & Error
2) Binomial Distributions
3) The Normal Distribution
4) Defining the Mean of a population

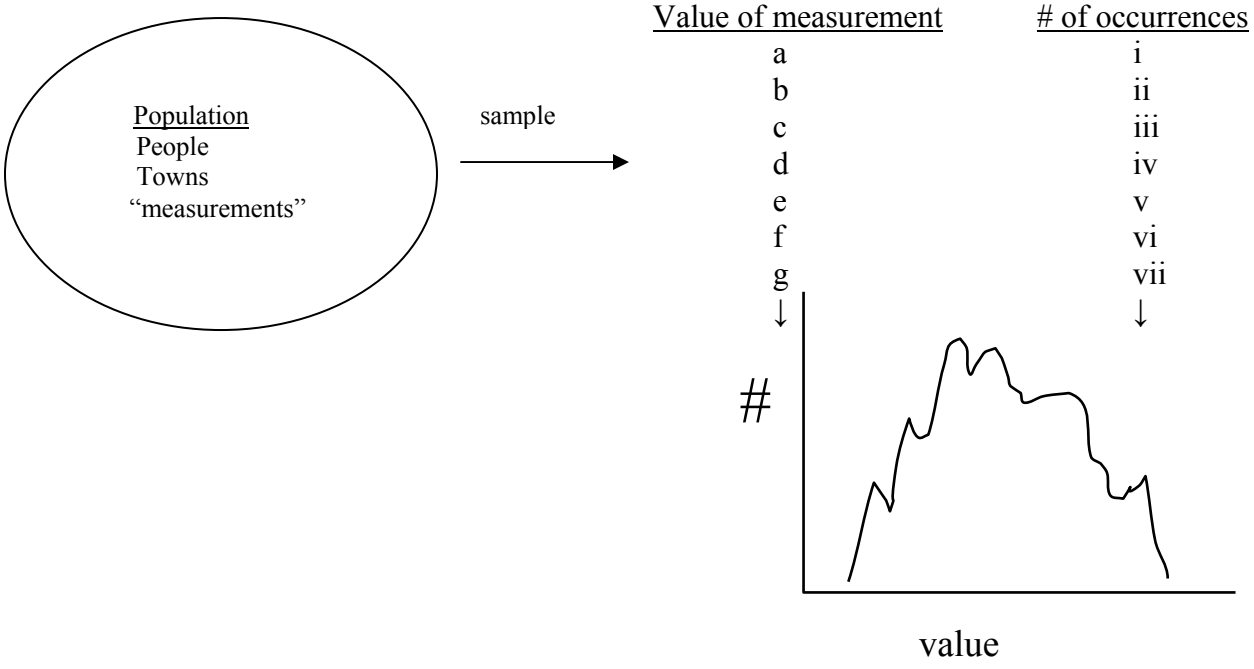
Goals: 1) Understand the concepts that underlie common statistical analyses for EHS studies
2) Evaluate them! Appropriate use? Quality? Meaning?
3) Requires familiarity with operations & mathematics; but not formal training to be a statitician

Why do we need statistics in EHS?

1) To organize data for analysis
2) To quantify error
3) To quantify and compare variation
4) To detect differences between populations
   e.g. affected vs non-affected
   exposed vs non-exposed
5) To detect relationships
6) To make predictions about events in the future; specifically to estimate risk

The Basic Tool for Organization of Population Data:

Frequency Plot or Distribution



1) All measurements and observations are samplings (e.g., heart rates, bead numbers, hand measurements)

There is an ideal universe of all our measurements that is infinite. We are tying to estimate the characteristic features of this ideal universe of measures.

Why? To develop the Best representation of reality. The Best representation of risk, prediction, difference.

If we measured body weight, how would the plot look?

We expect variability in this measure… its expected variation… remember this type of variance is always present (biological, quantitative, statistical)

Now, suppose I gave each of you a ruler (12 inches) and asked you to measure this table. How would the distribution of measurements look?

<Graphs>

What if I gave you 100 1-foot rulers and asked you to measure them with a 2-ft. ruler?

<graph>

We called the spread in data variance (mathematical definition later.)

Three sources of variance in population distributions:

Variance } I) Errors of measurement (quantitative, investigator, seasonal)
II) Variation (statistical, sampling, biological, physical)
III) "Things are really different. More than one distribution present."

I) Errors of Measurement

1) Technical: calibration, systematic
2) Investigator: ability, judgment, bias: more later for sure!
3) Process variation: temp-dependent, diurnal rhythms, seasonal
4) Population heterogeneity: samples not homogenous (w/ persons)

mixing is never complete

II) Variation

1) Known
   A) Controllable- diet, fasting, activity level
   B) Uncontrollable- gender, height, age (biological/physical)
      But can "control" for by matching strategies and adjustments (e.g., per capita, age-specific)
2) Unknown
   "Random;" "Statistical Variation" due to sampling from the universe of possibilities

III) Real Differences: What we seek to evaluate

Back to Distributions

How do we describe them?

How do we compare them?

Analysis of Distributions

1) We could use mathematics to develop an exact function to describe. HARD to do! Especially for each that is studied! f(v) = #

OR

2) We can model to idealized distributions for which mathematical treatments are tractable (though still not trivial!) - statistical methods

Three related distributions for statistical analyses:

Binomial- Based on frequency of occurrences when there are only two possible outcomes (e.g., bead drawings: each bead is either one color or the other; coin flipping: each toss is heads or tails). Well-developed mathematics, but two approximations are used for population data.

<Graph>

Poisson- For infrequent events

Normal- For frequent events

As sample size increases both Binomial & Poisson distributions converge to Normal.

Normal Distribution (ND)- Ideal
"Arises when the data (x values) are the sum of many independent small random factors" e.g. measurements, BP, heart rate

Also known as Gaussian Distribution, Bell Curve

<Graph>

Properties:
1) Symmetric
2) Most frequent value = mode = μ, mean, arithmetic average

$$\mu = \frac{\sum x_i n_i}{N}$$

3) mode = mean = median, value of x that accounts for 50% of the total values

Two parameters define the ND entirely

1) μ, the mean indicates where the distribution is centered
2) σ, the standard deviation $= \sqrt{\frac{\sum(\mu - x_i)^2}{N}}$

   Measure of the spread in the data around μ.

   E.g. For measurements, its an indicator of precision.

Now, how can we compare the spread of a distribution or the quality of precision at different scale and for measurements of different type.

For example: Measure of bacteria: scale = microns
Measure of temperature: scale = degrees
Measure of rat tails: scale = inches

CV, coefficient of variation = $(\sigma/\mu)\ 100\%$
μ

68% - 95% - 99% Rule- "As a probability density function"
$\pm 1\sigma$ $\pm 2\sigma$ $\pm 3\sigma$

Statistical methods based on the ND employ the two parameters σ and μ
Therefore they are called "**PARAMETRIC STATISTICS**"

Important Rule
Before you apply parametric statistics, confirm that the data are normally distributed. If not…

1) Transform to ND and then use parametrics (e.g., Log normal transformation).

2) Use non-parametrics (which are often conversions to "pseudo normal distributions, e.g. Mann Whitney)

What happens if you use parametric statistics when data are not normally distributed?

You may miss differences that more appropriate statistical methods might have sufficient power to detect.

What happens if you use non-parametric statistics when data are normally distributed?

You may miss differences that would have been detected by parametric methods; because when data are normally distributed, parametric statistics are the most powerful methods.

What type of formal statistical errors are these?

Move from ideal discussion to practical discussion

We want to know about μ in most cases

Consider:
An ideal population
With $N \rightarrow$ infinity (e.g., stars)
or N large, but finite (e.g., people)
and some measured property that is **normally distributed** about a population μ with standard deviation σ

When we measure a sample of n individuals,
we can construct a new sample distribution
with mean x-bar

x-bar, the sample mean is an estimate of what we seek, μ.

First Concern- How close is x-bar to μ?

Consider n = 1: x-bar could be near, could be far
x-bar is most likely to be near because of the normal probability density function, but it is not likely to = μ

Consider how x-bar approaches μ, as n approaches N

Central Limit Theory- "convergence to the mean": as the size of a sample increases, its mean, x-bar, approaches the mean, μ, of the sampled population

****

Now consider:
If we knew σ, the population standard deviation, (often we can, e.g., infant birth weights; heart rate) what can we say about how close x-bar, the sample mean, is likely to be near μ, the population mean?

Consider a single x again:
Given that the population is normally distributed,
[How could we tell? Evaluate the sample distributions form.]
we can say that our x is within $\pm 2\ \sigma$ of μ with 95% confidence.

<Graph> I.e., the 95% Confidence Interval (CI) for μ about $x = x \pm 1.96\ \sigma$

What does this mean? It means that, if we drew an x many times, 95% of the time μ, the mean of the sampled population mean, would lie within this interval. Therefore, we have set a limit on what μ might be.

What happens to this interval as we increase the sample size, the size of n?
Because of the central limit theory, our confidence that x-bar = μ increases.

Mathematically, the interval shrinks by the √n

And…

95% CI for μ about x-bar $= \text{x-bar} \pm \frac{1.96\,\sigma}{\sqrt{n}}$

Note that as n approaches N,

$\frac{1.96\,\sigma}{\sqrt{n}}$ approaches 0, and, therefore, x-bar approaches μ

For analysis of measurements, a

Small 95% CI implies precision

Large 95% CI implies uncertainty

How would you determine 99% CI?

$= \text{x-bar} \pm \frac{2.94\,\sigma}{\sqrt{n}}$

What does the 95% CI mean?

1) You are "95% confident that over many samplings, μ will lie in this interval"
2) If you performed 100 samplings of the same n, 95 of the x-bar's are predicted to lie in this interval and 5 will not. Why? Because the x-bar is an estimate of the population's mean, μ, which has a 95% probability of being within this interval.

****

Often, σ, population standard deviation is not known.

What then?

Then we can estimate σ from s, the sample standard deviation

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

n-1, Not n
When n approaches N, this is not as important, as in the calculation of σ.
It matters when the sample size is small.

n-1 = degrees of freedom

We use n-1 instead of n, because the x-bar is used in the calculation and x-bar is known.

Estimate of σ

CI for μ = x-bar $\pm \frac{ts}{\sqrt{n}}$

Note: as n approaches N,
t approaches 1.960 for 95% CI (see t-Table)
and of course
s approaches σ

t = *t*-statistic, a parameter that specifies the desired CI for a given n

So,

95%CI for μ = x-bar $\pm \frac{1.96\ \sigma}{\sqrt{n}}$

Look familiar?