BE.104 Spring
Biostatistics: Detecting Differences and Correlations
J. L. Sherley

Outline

1) Review Concepts
2) Detecting differences and quantifying confidence
3) Detecting relationships and quantifying confidence

Variance = σ^2

What do changes in variance tell us?
(Review in-class exercises)

<graphs>

Multiple "populations" present
Skewed data; non-normal data

An important distinction about the application of normal statistics that is often confused:

The sampled **POPULATION** should be normally distributed- why?

Question: If a sample distribution is not normal can we apply parametric statistical methods?

Yes, if they are a sample from an "ideal population" that is normally distributed. It is the properties of the ideal population that matter, not the distribution of the sample, per se. Caveat?

<graph>

Parametric statistical methods address the uncertainty of sampling.

Now we focus on the structure of the sample because we know it gives us some information about the structure of the ideal population.

So, we must base our decision about using normal statistics on the sample when we have no a priori information about the structure of the ideal population with N members.

Now to detecting differences

What we want to ask is:

Are two means more different than we would expect based on "error" & statistical variation alone?

Consider there is only one ideal population, and we may be looking at sampling variation or statistical variation and error.

<picture>

So, we define the possible range of differences in the means that could occur by chance/error with some level of confidence.

<graph>

If this difference is not explained by error & statistical variation (i.e., variance) we then can consider other factors:

E.g., A change in a physiological mechanism

Which might lead us to:
Parallel testing (i.e., a bigger study)
Orthogonal testing (i.e. different kind of study; intervention experiment)
"More on this later"

"Detecting Differences Between Estimated Pop. Means"

Our question can be phrased this way regarding the ideal population thinking:

<picture with graph>

Two sample distributions, with numerical different sample means, drawn from ideal populations A and B. Hypothesis: A and B are distinct populations with distinct population means, μ.
The null hypothesis: the observed numerical difference occurs due to variance, and the two sample distributions actually derive from the same population.

What do we need to consider for this evaluation?

Given:

1) the magnitude of $\bar{x}_A - \bar{x}_B$
2) the variance about $\bar{x}_A + \bar{x}_B$
3) Variance = $\sigma^2 \approx s^2$

Different ways to state the question:

> What is the probability that you will call them different when they are not?
>
> What is the probability of being wrong when you think that Pop_A and Pop_B are different?
>
> What is the probability that the observed numerical difference in sample means occurs due to chance & errors when population means are equal ($\mu_A = \mu_B$); i.e., there is really only one population that the sample is drawn from?

William Sealy Gosset developed a statistic that gives this probability.
Published in 1908 in *Biometrika* under the pseudonym "Student"

Student's t-statistic (originally called "z")

$$t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{(s_A^2/n_A) + (s_B^2/n_B)}} = \frac{\text{difference in sample means}}{\text{standard error of the difference in the sample means}}$$

t defines the probability that the observed numerical $\bar{x}_a - \bar{x}_b$ occurs by chance
t defines the probability that they are equal when you think that they differ
t defines the probability of being wrong when you think the null hypothesis is not supported

The distribution defined by t is a normal distribution

<Develop graphical explanation>

t – distribution properties

1) At a given *t*, two ways to be wrong ⇒ "2-tailed test"
   μ_A not > μ_B; and μ_A not < μ_B

   One tail: μ_A not ≥ μ_B only; or μ_A not ≤ μ_B only
   E.g., measurements that cannot be negative versus zero: $\mu_0 \leq \mu_B$ only
   BE WARY- the one tail statistic gives the same level of significance at 1/2 the power (more later)

2) As *t* increases, the probability of being wrong, when you think $\mu_A \neq \mu_B$, decreases
   I.e., large *t* values are GOOD for the hypothesis that there is a difference in the means.

3) As n increases, *t* increases
   Tables allow us to assign confidence levels (p values) for *t* @ a given n

See *t*- table; Schork and Remington A-5

p is inversely related to *t* (p = 1- level of confidence that the null hypothesis is not supported)
I.e., small p is GOOD for the hypothesis that there is a difference in the means.

$p < 0.05$

Predicts that if you performed this comparison (x bar_A vs x bar_B) 20 times, you would be wrong about their being different ≤ 1 time, when you thought they were from different populations.

"< a 5% probability that an observed numerical difference occurs due to chance/error"

What is the specific method?

1) Compute *t* from x bar_A, s_A, n_A and x bar_B, s_B, n_B

2) Go to t-statistic table: df = $n_A + n_B - 2$, indicator of sample size

3) Extract p, probability of being wrong, when you conclude that two samples came from different populations.

Other ways to think about the *t*- stat

<Graph>

1) Suppose we computed the 95% CI for μ_A and μ_B from x bar_A, s_A and x bar_B, s_B?

What can we say if the intervals don't overlap?

What can we say if the intervals do overlap?

2) x bar_A- x bar_B is an estimate of the "Population of μ_A - μ_B values."

Therefore, we can compute 95% confident interval for μ_A - μ_B about x bar_A- x bar_B

95% CI for μ_A - μ_B = (x bar_A- x bar_B) ± ($t_{0.05}$) ($\sqrt{(S^2_A/n_A) + (S^2_B/n_B)}$)

What would you conclude if the 95% CI for μ_A - μ_B about
x bar_A- x bar_B = 2.5 was -0.5 to 5.5 ?

Sometimes we are comparing a data set to a single value.
E.g., How does a set of measurements compare to a standard value, sv (not SD and variance!)

<drawing>

$$t = \frac{sv - \text{x bar}}{s/\sqrt{n}}$$

95% CI for μ-x = (x bar – sv) ± $t_{0.05}$(s/√n)

Similar: determine 95% CI for μ
If sv outside of interval? "Outlier Evaluation"
If sv in interval?

Paired t-test (versus "unpaired")

Sometimes two sets of data are paired (e.g., heart rate before & after Rx)
Advantage- Avoid intra-set variation due to known causes or effects
e.g. day-to-day variability (before vs after treatment); lot to lot variation

<drawing> Calculate differences d, d bar, and s_d from the distributions: t_d = (d bar)/(s_d/√n)

Caution- Must be paired a priori to avoid bias! (same time, conditions, etc.)
Use when s_1 and s_2 are large. When s_1 and s_2 are already small, pairing reduces the sensitivity of the test. Why?

Detecting & Quantifying Associations

Encounter lots of potential causative factors, where

Factor X → Disease/Toxicity

How do we begin to move from conjecture and anecdotes to causes and effect?

Quantify Associations- Place a quantitative value on how much Factor X is associated with changes in disease/toxicity

1) Concordance analysis- "X → disease/toxicity?"; and
"Not X → Not disease/toxicity?"

More later on this.

2) OR, RR ideas developed earlier

3) Correlation- How does disease/toxicity vary with changes in X?
1) similar to "dose response"
2) location, time, persons, factors, too!

<graph>

General Discussion for Detecting Associations or Correlations

Begin with consideration of a population of related variables X and Y, with Y dependent on X suspected. (e.g. diet vs weight).

< X vs Y scatter plot>

We call Y, the dependent variable; and X, the independent variable

The typical problem we confront:

<diagram>

Question: What is the nature of the population from which the sample was drawn?

1) Was there an association between X and Y?

OR

2) Does the observed association occur by error/chance in the sampling?

First Approach

Determine best fit line to the data by linear regression. "Least squares fit"

<Graph>

$\Sigma (Y_i$ - line Y-value)^2

Minimize $\Rightarrow$ regression line: $y\hat{} = a + bx$

b, the slope, is a measure of the degree to which Y depends on X

When b = 0, Y does not depend on X
b > 0, Y is positively correlated with X
b < 0, Y is negatively correlated with X

Can detect associations with 1st approach, but no quantification of confidence that the data are not so arranged by chance. Our question is not, "How well does the regression line fit the data?"(as e.g., the root mean squared deviation would be).

It's, "How likely is it that the observed association occurred by chance/error?"

Second Approach

The linear regression approach assumes that dependency is known.
I.e., in our example Y depends on X.
However, the regression result can be different if the dependency is reversed:

Y vs $y\hat{}$ variability $\neq$ X vs $x\hat{}$ in some cases

Pearson Product- Moment Correlation Coefficient

1) Asks: Is there an association when dependency relationship is not established?
2) Gives a quantitative measure of the strength of a detected association
3) Gives a measure of the confidence that a detected association occurs for reasons other than chance/error

$$r = \frac{\Sigma(X-Xbar)\,(Y-Ybar)}{\sqrt{\Sigma(X-Xbar)^2\,\Sigma(Y-Ybar)^2}}$$

$-1 \leq r \leq 1$

When r = 0, no association
r = 1, strong positive correlation

i.e., 100% of the variation in one variable is accounted for variation in the other

r = -1, strong negative correlation

How does r relate to the regression analysis?

$$r = \sqrt{1 - \frac{\sum [Y-(a+bx)]^2}{\sum (Y-Ybar)^2}}$$

$V_{residual}$, variance about the regression line
V_{tot}, , variance in Y data

$$r = \sqrt{1 - \frac{V_{res}}{V_{tot}}}$$

1) As $V_{res} \rightarrow 0$, there is no variation about the regression line
$r \rightarrow 1.0 \Rightarrow$ Y predicted by X w/o uncertainty

2) When $V_{res} \rightarrow V_{tot}$, $r \rightarrow 0 \Rightarrow$ no trend in data: X cannot be used to predict Y

$$r^2 = 1 - \frac{V_{res}}{V_{tot}}$$

"coefficient of determination"
"R^2" sometimes

$$0 \leq r^2 \leq 1$$

"Gives an intuitive feel for how well a straight line describes relationship between X and Y."
The degree to which Y depends on X.

Statistics have been developed for a given r or r^2 and $n_x + n_y - 2$ degrees of freedom to determine p value for confidence that the association between x + y does not occur by chance/error.
[Not: "How well does the line fit the data?"]