BIOSTATS 640 – Intermediate Biostatistics
Frequently Asked Questions
Topic 1 FAQ 1 – Review of BIOSTATS 540 Introductory Biostatistics

---

1.
I'm confused about the jargon and notation, especially population versus sample. Could you please clarify?

In BIOSTATS 540, we were introduced to the notion of there being a population in the background: a source population, about which we would like to learn). We were also introduced to the idea of a sample drawn from that population: a collection of actual, observed and known, data values that we will use to draw some inferences about the population. We were reminded that, in real life, typically we do not have the luxury of examining the entirety of a population (that would be a census).

As regards jargon and notation, the convention is to use greek letters to represent characteristics of the source population (and we referred to these as parameters and parameter values) and roman letters to represent characteristics of the sample. Reminder: a *statistic* is just a number that you calculate from the data in a sample.

So here is a little refresher schematic that we might have compiled in BE540 so as to keep track -

	Parameter in Population	Estimate from Sample
Mean	μ	$\bar{X}$
Variance	σ^2	S^2
etc …		

Here in BIOSTATS 640, when we learn about regression and correlation, a similar compilation allows us to keep track of what's what. Keep in mind that, typically, the statistic we calculate is calculated as our guess of the parameter in the population.

	Parameter in Population	Estimate from Sample
Slope of line of Y on X	β_1	$\hat{\beta}_1$ or b_1
Intercept of line of Y on X	β_0	$\hat{\beta}_0$ or b_0

Additional note to reader: See the little hat on top? Whenever you see the little hat on top, this is telling you that what you are looking at is an estimate. It's unfortunate that the letter is greek but the key is to notice that the little hat means the quantity is an estimate obtained from the data and is therefore a statistic. The little hat also goes by the name ***"caret".*** *Whenever you see it, think estimate.*

2.
Remind me again of the distinction between standard deviation (SD or S) and standard error (SEM or SE) and how this is related to the distinction between populations versus sampling distributions.

SD or S - Standard deviation is the $\sqrt{\text{variance}}$ of values of individuals in nature.

SEM or SE - Standard error is the $\sqrt{\text{variance}}$ of values of a statistic.

The collection of all possible individuals in nature goes by the name **"population"**
The collection of all possible values of a statistic (imagine replicating your study over and over a gazillion times and compiling the collection of all possible sample means $\bar{X}$) does not go by the name "population", even though this would make sense. Instead, this collection of all possible values of whatever statistic you're interested in goes by the name **"sampling distribution"**.

So who cares? Well, actually there are times when we are very interested in the sampling distribution of $\bar{X}$ (eg – clinical trials). And there are times when we might be interested in the sampling distribution of S^2 (eg – studies of lab performance). By extension, we can imagine that there might be times when we're interested in some other statistic.

In our unit on regression and correlation for example, we will see that we are interested in the sampling distribution of an estimated slope, $\hat{\beta}_1$

" I guess I don't see why I'd be interested in the sampling distribution of $\hat{\beta}_1$ *"*

You're interested in the sampling distribution of $\hat{\beta}_1$ when you're interested in what another investigator might obtain as a $\hat{\beta}_1$ if he/she were to repeat your study and come up with his/her own estimate. Whether you're aware of this consciously or not, this is the kind of thing you are interested in (generalizability, robustness are some familiar terms for this) when you read a journal article and are wanting to know if you would obtain similar findings if you were to repeat the published study in your own sample of folks.

3.
Ick. I don't understand summation notation.

Unfortunately, notation does get in the way of understanding ideas sometimes. The summation notation is nothing more than a secretarial convenience. We use it to avoid having to write out long expressions. For example,

Instead of writing $x_1 + x_2 + x_3 + x_4 + x_5$,

We write $\sum_{i=1}^{5} x_i$

Another example –

Instead of writing $x_1 * x_2 * x_3 * x_4 * x_5$,

We write $\prod_{i=1}^{5} x_i$

This is actually an example of the product notation

Key to the summation notation

$\sum$ The Greek symbol sigma says "add up some items"

$\sum_{\text{STARTING HERE}}$ Below the sigma symbol is the starting point

$\overset{\text{END}}{\sum}$ Up on top is the ending point

4.
What are Z-scores, what are t-scores and what is the distinction between them?

The **Z-Score** is a tool to compute probabilities of intervals of values for X distributed Normal(μ,σ^2).

Suppose it is of interest to calculate a probability for a random variable X that is distributed Normal(μ, σ^2). Sometimes (less so as time goes on because internet resources are getting better all the time), we're in a pickle because tabulated normal probabilities are available only for the Normal Distribution with $\mu = 0$ and $\sigma^2=1$. We solve our problem by exploiting an equivalence argument. The technique goes by the name **"standardization"**and involves replacing the desired calculation with an equivalent one for a new random variable called a **z-score.**

"Standardization" expresses the desired calculation for X distributed Normal(μ,σ^2) as an equivalent calculation for Z (Z is now called a Z-score) where Z is distributed standard normal, Normal(0,1).

$$\text{pr}\left[a \le X \le b\right]=\text{pr}\left[\left(\frac{a-\mu}{\sigma}\right)\le Z \le \left(\frac{b-\mu}{\sigma}\right)\right].$$ Thus,

$$\text{Z-score} = \frac{X-\mu}{\sigma}$$

- *Note* - The technique of *standardization* of X involves *"centering"* (by subtraction of the mean of X which is μ) followed by *"rescaling"* (using the multiplier $1/\sigma$)

Watch out when you are performing standardization that the re-scaling is with the correct $\sqrt{\text{variance}}$. Here are 3 examples followed by a generic, just to be sure that you get the idea:

1. $$\text{pr}\left[a \le X \le b\right]=\text{pr}\left[\left(\frac{a-\mu}{\sigma}\right)\le \text{Z-score} \le \left(\frac{b-\mu}{\sigma}\right)\right]$$

2. $$\text{pr}\left[a \le \bar{X}_n \le b\right]=\text{pr}\left[\left(\frac{a-\mu}{\sigma/\sqrt{n}}\right)\le \text{Z-score} \le \left(\frac{b-\mu}{\sigma/\sqrt{n}}\right)\right]$$

3. $$\text{pr}\left[a \le \hat{\beta}_1 \le b\right]=\text{pr}\left[\left(\frac{a-E(\hat{\beta}_1)}{SE(\hat{\beta}_1)}\right)\le \text{Z-score} \le \left(\frac{b-E(\hat{\beta}_1)}{SE(\hat{\beta}_1)}\right)\right]$$

4. $$\text{pr}\left[a \le \text{statistic} \le b\right]=\text{pr}\left[\left(\frac{a-E(\text{statistic})}{SE(\text{statistic})}\right)\le \text{Z-score} \le \left(\frac{b-E(\text{statistic})}{SE(\text{statistic})}\right)\right]$$

The z-score method is appropriate under 2 circumstances: (1) when the starting variable is distributed Normal to begin with, and (2) when the starting variable can be appreciated as an instance of the central limit theorem (not discussed here).

A **t-score** is a student's t random variable. There's lots of ways to have a random variable that is distributed student's t. One is to conceive of a student's t random variable as a t-score and in this way, analogous to a z-score.

One definition of a student's t random variable: In the setting of a random sample $X_1 \ldots X_n$ of independent, identically distributed outcomes of a Normal(μ, $\sigma 2$) distribution, where we calculate $\bar{X}$ and S^2 in the usual way:

$$\bar{X}=\frac{\sum_{i=1}^{n} X_i}{n} \quad \text{and} \quad S^2=\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{n-1}$$

a student's t distributed random variable results if we construct a t-score instead of a z-score.

$$t\text{-score}=t_{DF=n-1}=\frac{\bar{X}-\mu}{s/\sqrt{n}}$$ is distributed Student's t with degrees of freedom = (n-1)

Note – If we want to standardize $\bar{X}$, the solution depends on whether we know its SE or we don't.

	$SE(\bar{X})$ is known	$SE(\bar{X})$ is NOT known
Standardization of $\bar{X}$	$Z\text{-score}=\frac{\bar{X}-\mu}{SE(\bar{X})}$	$t\text{-score}=\frac{\bar{X}-\mu}{\hat{SE}(\bar{X})}$
Where	$SE(\bar{X})=\frac{\sigma}{\sqrt{n}}$	$\hat{SE}(\bar{X})=\frac{S}{\sqrt{n}}$

Recall $S^2=\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{(n-1)}$