

Unit 1. Review of Introductory Biostatistics

“In solving a problem of this sort, the grand thing is to be able to reason backwards!”

- Sherlock Holmes

Nature is inherently variable. Measure a person’s blood pressure twice and you will get different results. Measure the blood pressures of different individuals and you will get different results. We care that nature is variable when we have important questions to answer. Is a new treatment significantly better than standard care? Does a particular environmental exposure pose a significant health risk? Is a parent’s genetic profile a significant predictor of birth outcome? What do we mean by *significant* anyway? In statistical parlance, a statistically significant result is one that was unlikely to have occurred by chance. Couple (1) “*unlikely to have occurred by chance*” with (2) a *large effect* (“*effect size*”) in (3) a *large sample size* (“*study design*”) and you are on your way to answering your question. The ideas and methods of biostatistics provide numerical tools for assessing chance, effect size, and meaningful inference. PubHlth 540, *Introductory Biostatistics*, delivered many tools and emphasized statistical literacy:

- (1) Choosing a correct analysis depends on the type of data you have and the specific question you are trying to answer;
- (2) A continuous type variable can have any value on a continuum of real numbers (eg – blood pressure), whereas discrete type variable values are separated by gaps, the latter meaning that, between two values (eg “1 visit” versus “2” visits”, nothing in-between is possible;
- (3) For a 95% *confidence interval (CI)*, we are 95% confident that our random sampling captured the unknown true population parameter value (which is not random); and
- (4) Assessing the likelihood of a particular result (or a result more extreme), under a presumed null hypothesis (*p-value*), is done using probability theory.

Nature

Population/
SampleObservation/
DataRelationships/
ModelingAnalysis/
Synthesis

Table of Contents

	Topic	Page
1	Signal to Noise	2
2	Description and Estimation	7
3	Just so it's clear – SD versus SE	13
4	The Sample Average and the Central Limit Theorem	15
5	The Statistical Confidence Interval	18
6	Statistical Hypothesis Testing	24

1. Signal to Noise

In *PubHlth 540 Introductory Biostatistics*, we learned that if a ratio $\frac{\text{signal}}{\text{noise}}$ is large and unlikely, we say it is “significant”

The variations that are likely to occur are collectively called “*noise*”. For example, all other things being equal, my blood pressure tomorrow (eg., 105/70 mm Hg) will be somewhat different than my blood pressure today (eg., 110/65 mm Hg), and I will interpret this as just noise. A “*signal*”, on the other hand, is a large variation that is unlikely to have occurred. For example, if my blood pressure tomorrow is as different as 160/100 mm Hg, I will interpret it as both very different from 110/65 mm Hg and unlikely. I will interpret this as a “signal” that something is significantly different.

Significance is relative. How large a difference does a difference need to be to matter? Without additional analysis, we don't know.

- **Example** - In 1969, the average number of serious accidents per 100,000 workers/year was 10. In 2011, the average was 7.

Is the downward trend real?

- **Example** - Voyager 2 is circling Saturn. A “blip” appears on our receiver.

Is it signal or noise?

We used the idea of “signal-to-noise” in estimation and hypothesis testing.

Signal: Treatment effect, exposure effect, secular trend

Noise: Natural variation, random error, chance

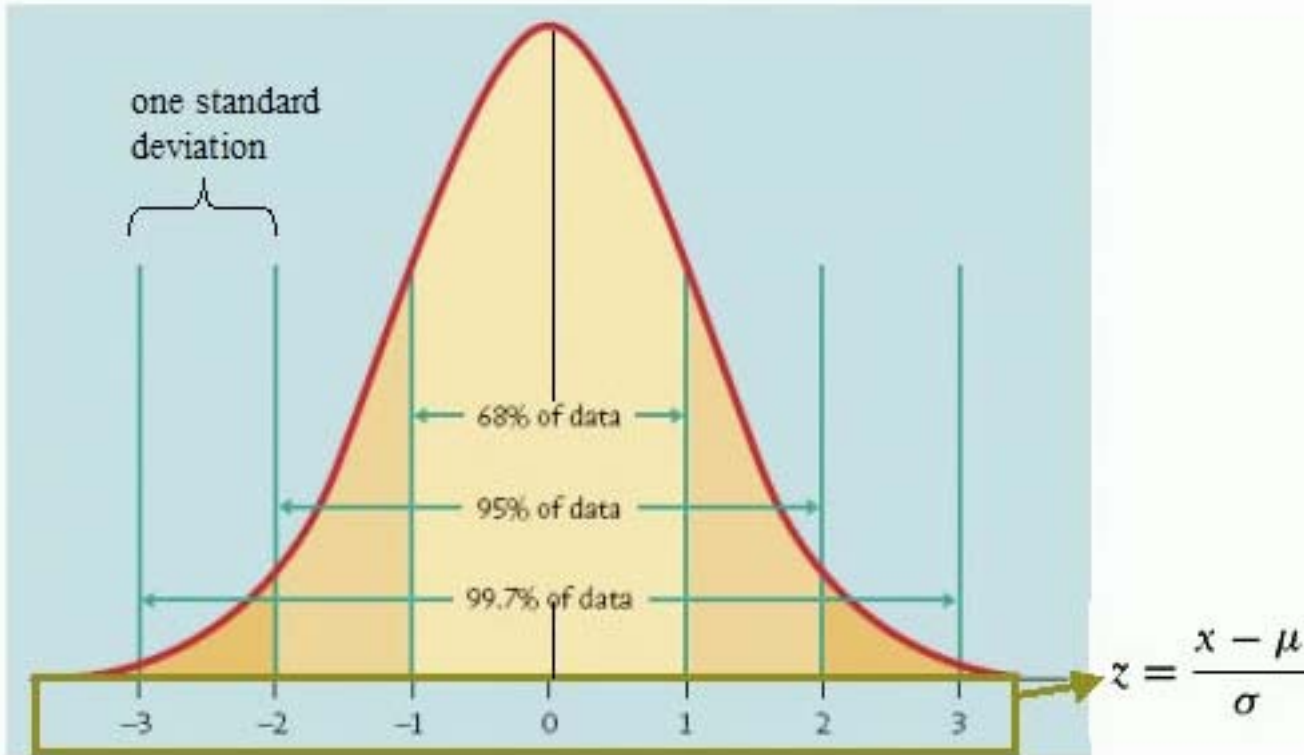
Nature ——— Population/
Sample ——— Observation/
Data ——— Relationships/
Modeling ——— Analysis/
Synthesis

We learned that a Z-score (and similarly a t-score) = $\frac{\text{observed} - \text{expected}|\text{null}}{\text{se}(\text{observed})} = \frac{\text{signal}}{\text{noise}}$.

This is great, because Z-scores are distributed standard normal and we can use the normal distribution to define “benchmarks” for discriminating “signals” from “noise”.

Plot of the Standard Normal Distribution

(source: Google images)



The standard distribution is centered at its mean value $\mu=0$.

Along the horizontal axis are tick marks at increments of 1 standard deviation $\sigma=1$

Because 68% of the distribution is captured by z-score values in the range (-1, +1), we might interpret observations that are within ± 1 sd of the mean as “noise”.

We might also say that, because 95% of the distribution is captured by z-score values in the range (-2, +2), then data values within ± 2 sd of the mean are also “noise”.

However, when we see z-score values that are as extreme as (or more extreme than) ± 3 sd in distance from the mean, we interpret them as large and unlikely. These are “signals” that warrant further investigation.

Nature

Population/
Sample

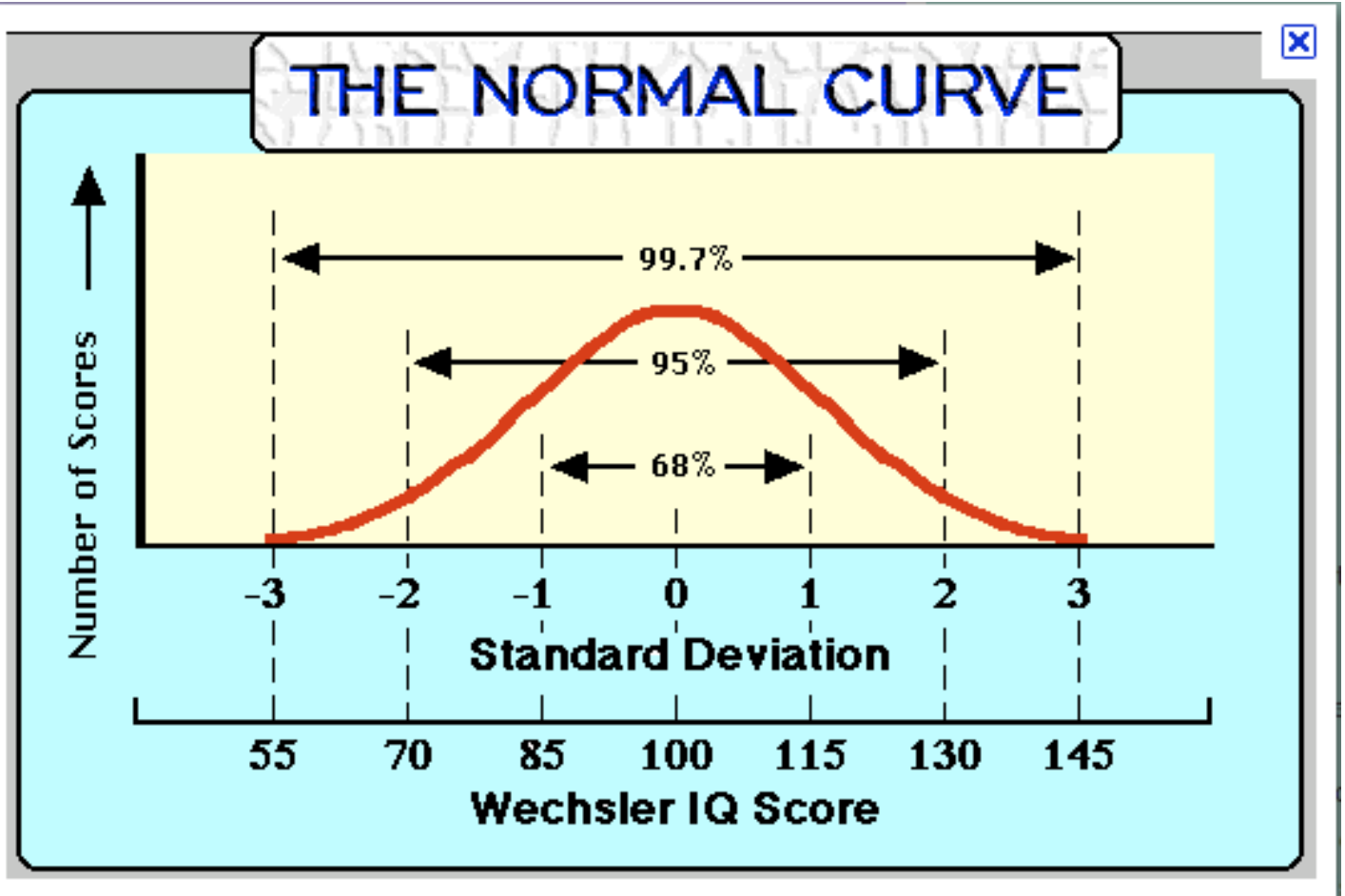
Observation/
Data

Relationships/
Modeling

Analysis/
Synthesis

Example - The Wechsler IQ Score Distribution

(source: Google images)



The distribution of Wechsler IQ scores has mean value $\mu = 100$ and standard deviation $\sigma=15$. Thus, “sd” tick marks on the horizontal axis are in increments of 15 Wechsler IQ score points.

68% of the distribution of Wechsler IQ scores are in the range (85, 115).

95% of the distribution of Wechsler IQ scores are in the range (70, 130).

IQ scores below 70 or above 130 are unlikely as they occur in 5% or less of the population. We might interpret these as “signals” warranting further study.

Nature

Population/
Sample

Observation/
Data

Relationships/
Modeling

Analysis/
Synthesis

A cautionary note. Signal-to-noise is a matter of perspective.

Consider the following hypothetical situation.

A randomized controlled trial of a new treatment for advanced melanoma is not statistically significant. Patients receiving the new treatment survived longer, on average, but the finding did not achieve statistical significance.

- From a ***public health perspective***, the variability in outcomes in the experimental versus control groups reflect noise; whereas,
- For the ***individual patient***, any extra life afforded by the new treatment is hugely significant (signal).

2. Description and Estimation

Description. In *PubHlth 540, Introductory Biostatistics*, we learned how to extract and display meaningful summaries of the facts contained in a sample of data (graphs, tables).

We learned that the appropriate methods for description are different, depending on the data type.

Data Type				
Type	Qualitative		Quantitative	
	Nominal	Ordinal	Discrete	Continuous
Example	religion	Strength of opinion	# visits to doctor	weight
Descriptive Methods	Bar chart Pie chart - -	Bar chart Pie chart - -	Bar chart Pie chart Dot diagram Scatter plot (2 variables) Stem-Leaf Histogram Box Plot Quantile-Quantile Plot	- - Dot diagram Scatter plot (2 vars) Stem-Leaf Histogram Box Plot Quantile-Quantile Plot
Numerical Summaries	Frequency Relative Frequency Frequency	Frequency Relative Frequency Cumulative Frequency Frequency	Frequency Relative Frequency Cumulative Frequency means, variances, percentiles	- - - means, variances, percentiles

Source: *PubHlth 540 Lecture Notes 1 (Summarizing Data)*, page 11

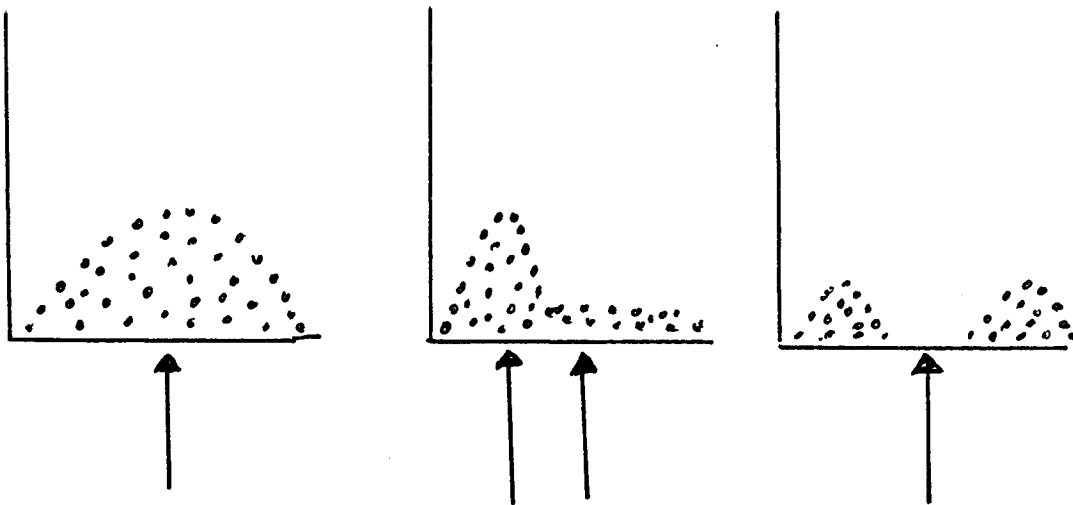


For continuous variable data, we learned that there is more than one choice for describing location and that these differ in their meaning and appropriateness.

- Arithmetic average (*mean*)
- Middle most (*median*)
- “Benchmarks” (*percentiles*)

Ditto for describing dispersion

- *Standard deviation (SD)*
- *Median of absolute deviation from the median (MADM)*



We learned the distinction between sample variance (S²) and sample standard deviation (SD) and how to do the calculation:

Sample variance (S²) S² is a summary measure of the squares of individual departures from the sample mean in a sample. Recall. We talked about this as an “almost average (because we divided by n-1)” distance from the sample mean.

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

Standard Deviation (S or SD). S (or SD) is the square root of S². The advantage of the square root operation is that the resulting summary has the same scale as the original values.

$$\text{Sample Standard Deviation (S or SD)} = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}}$$

Illustration of Calculation

Source: PubHlth 540 Lecture Notes 1 (Summarizing Data), page 43

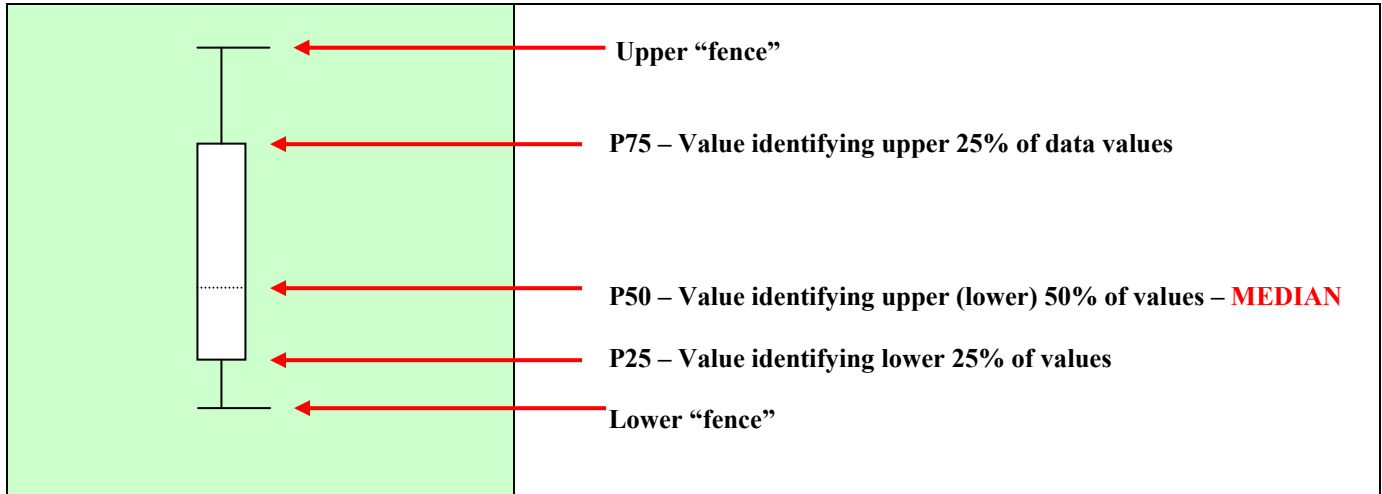
Patient Identifier, “i”	Survival (days), X _i	Mean for sample, \bar{X}	Deviation , (X _i - \bar{X})	Squared deviation (X _i - \bar{X}) ²
1	135	161	-26	676
2	43	161	-118	13924
3	379	161	218	47524
4	32	161	-129	16641
5	47	161	-114	12996
6	228	161	67	4489
7	562	161	401	160801
8	49	161	-112	12544
9	59	161	-102	10404
10	147	161	-14	196
11	90	161	-71	5041
TOTAL	1771		0	285236

◆ $\sum_{i=1}^{11} X_i = 1771 \text{ days} \rightarrow \bar{X} = \frac{1771}{11} = 161 \text{ days}$



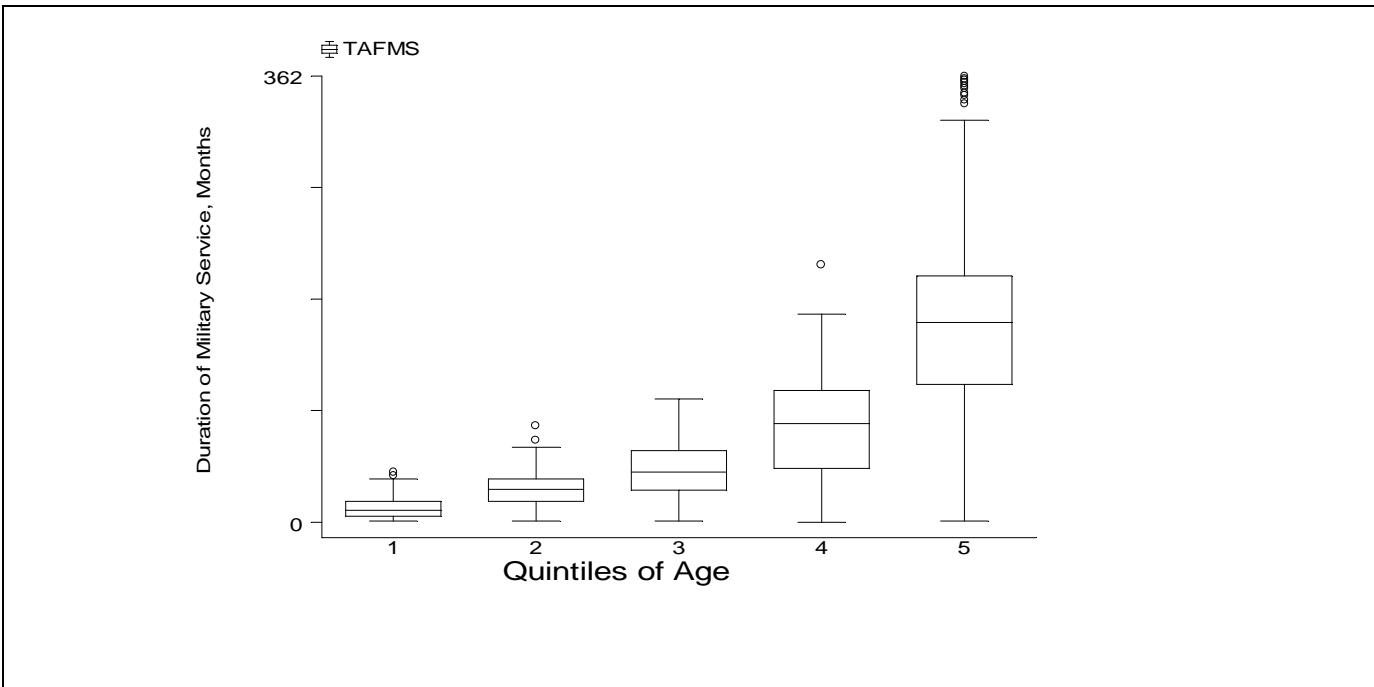
- ◆ **Sample variance is** $s^2 = \frac{\sum_{i=1}^{11} (X_i - \bar{X})^2}{n-1} = \frac{285236}{10} = 28523.6 \text{ days}^2$
- ◆ **Sample standard deviation is** $s = \sqrt{s^2} = \sqrt{28523.6} = 168.89 \text{ days}$

The **box plot summary** took a little getting used to.



- The central box has P₂₅ and P₇₅ for its limits. It spans the middle half of the data.
- The line within the box identifies the median, P₅₀. Sometimes, an asterisk within the box is shown. It is the mean. The lines coming out of the box are called the “whiskers”. The ends of these “whiskers” are called “fences”.
- Upper “fence” = The largest value that is less than or equal to $P_{75} + 1.5*(P_{75} - P_{25})$.
- Lower “fence” = The smallest value that is greater than or equal to $P_{25} - 1.5*(P_{75} - P_{25})$.

Illustration of Interpretation of a Box Plot – Duration of Military Service by Age

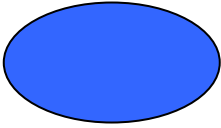

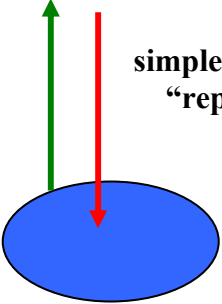


- Duration of military service increases with age.
- With age, the variability in duration of military service is greater.
- The individual circles represent extreme values.

Source: PubHlth 540 Lecture Notes 1 (Summarizing Data), page 51

Estimation. We also learned about random sampling and the role it plays in estimating values of population parameters.

When it can be assumed that the sample is representative of the population from which it was drawn, the values of the summary statistics computed from the sample can also be interpreted as estimates of the corresponding population parameter values.

Description of a Sample	Estimation
<p data-bbox="103 1039 207 1073">Sample</p>  <p data-bbox="103 1150 792 1180"><i>In <u>description</u>, there is no larger context than the available data.</i></p>	<p data-bbox="824 674 971 707">population</p>  <p data-bbox="1166 856 1479 926">simple random sample “representative”</p> <p data-bbox="824 1039 922 1073">sample</p>  <p data-bbox="824 1150 1511 1297"><i>In estimation, the advantage of the sample being the result of simple random sampling (red arrow) is that it permits inference backwards (green arrow) to the source population. The sample statistics are then estimates of the corresponding population parameters..</i></p>

3. Just so it's clear – SD versus SE

The **standard deviation (SD or S)** is a summary of the variability of **individuals in nature**, whereas

the **standard error (SE)** is a summary of the variability of a **summary statistic among many replications of your study**. Recall. We imagine a collection of values of a sample statistic such as the sample mean that is obtained by repeating your whole study over and over again.

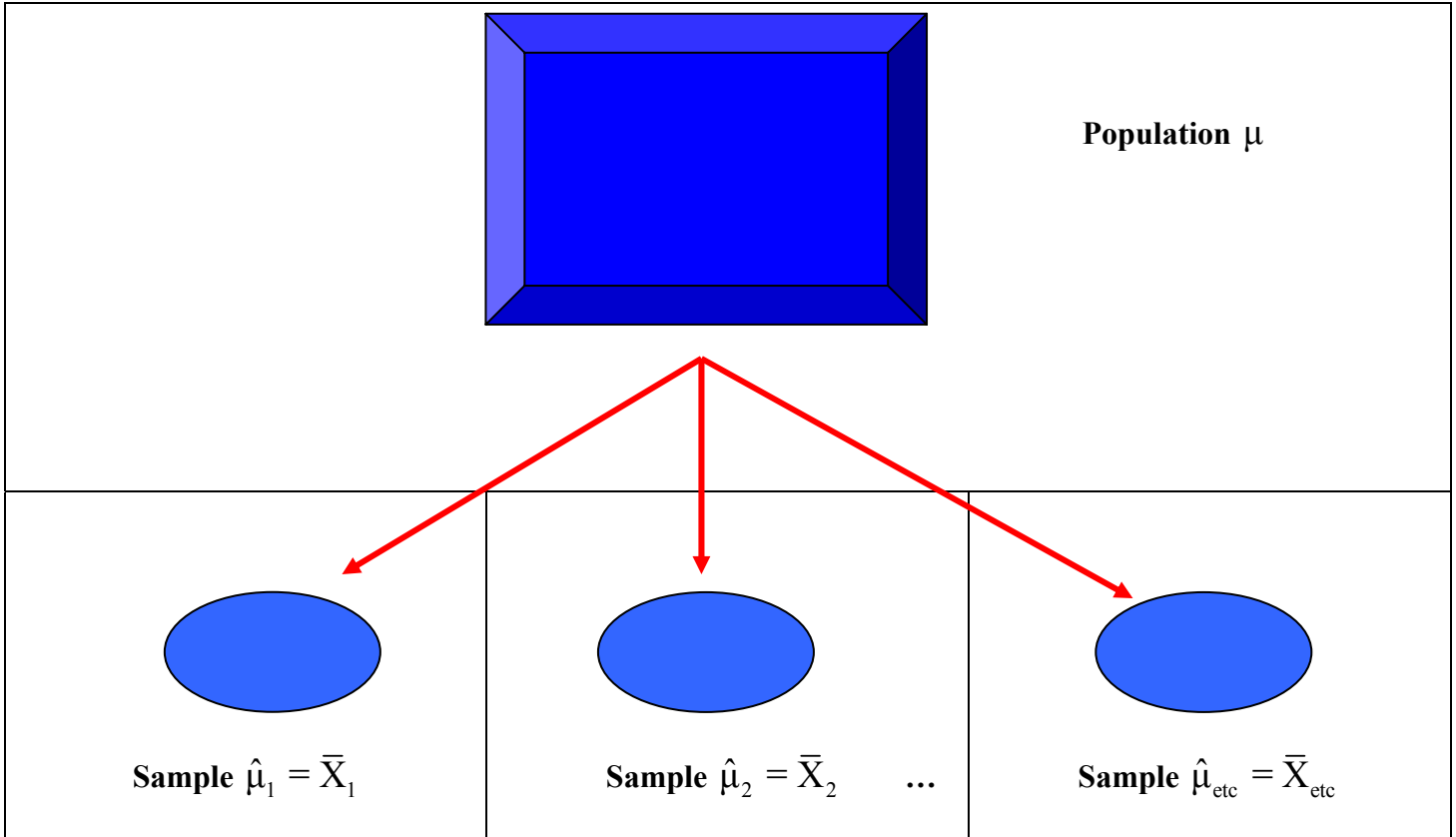
We clarified SD versus a particular SE, namely: the SE of the mean \bar{X}

<p><u>Standard Deviation</u></p> <ul style="list-style-type: none"> Standard deviation is a measure of the variation among the values of a measurement (eg – cholesterol) that are obtained on <i>individuals</i>. The standard deviation of the values of a measurement (eg – cholesterol) in a <i>population</i> is represented by the symbol σ The standard deviation among the values of a measurement (eg – cholesterol) in a <i>sample</i> is represented by either or two symbols: $\hat{\sigma}$ or S 	<p><u>Standard Error of the Mean \bar{X}</u></p> <ul style="list-style-type: none"> The standard error is a measure of the variation of among the values of a statistics (eg – sample mean \bar{X}) in the sampling distribution of that statistic (eg – the sampling distribution of \bar{X}). The standard error of a statistic is represented by the letters SE with a notation that identifies the statistic. For example, the standard error of \bar{X} is represented by the notation $SE(\bar{X})$. <p><i>Example -</i></p> <ul style="list-style-type: none"> Under simple random sampling from a population with mean μ and variance σ^2, $SE(\bar{X}) = \frac{\sigma}{\sqrt{n}} \text{ where } n = \text{size of sample.}$ <ul style="list-style-type: none"> When σ^2 is not known, our estimate is $SE\hat{E}(\bar{X}) = \frac{s}{\sqrt{n}}$
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------



We developed an appreciation for sampling distributions.

Example – A sampling distribution of the sample mean is a theoretical distribution that is obtained by replicating the process of simple random sampling over and over again, infinitely many times. With each replication of the sampling process, a new \bar{X} is obtained. The collection of these is analogous to a “population” but we don’t call it a population. Instead, we call it the sampling distribution of \bar{X} .



4. The Sample Average and the Central Limit Theorem *A Great Result!*

We learned that a lot of statistical inference is based on the normal distribution and that it is appropriate for continuous random variables only.

- The pattern of occurrence of many phenomena in nature happens to be described well using a normal distribution model.
- Even when the phenomena in a sample distribution are not described well by the normal distribution, the sampling distribution of sample averages obtained by repeated sampling from the parent distribution is often described well by the normal distribution (*Central limit theory*).

Normal Distribution (μ, σ^2)

A random variable X that is distributed normal with mean= μ and variance= σ^2 has probability density function

$$f_X(X=x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right] \text{ where}$$

x = Value of X

Range of possible values of X: $-\infty$ to $+\infty$

Exp = e = Euler's constant = 2.71828

π = mathematical constant = 3.14

μ = Expected value of X

σ^2 = Variance of X, which is the expected value of $[X - \mu]^2$

Standard Normal Distribution ($\mu=0, \sigma^2=1$)

A random variable Z that is distributed standard normal has probability density function

$$f_Z(Z=z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z^2}{2}\right]$$

The Central Limit Theorem

IF

- 1) We have an independent random sample of n observations $X_1 \dots X_n$; and
- 2) the $X_1 \dots X_n$ are all from the same distribution, *whatever that is*; and
- 3) this distribution has mean $= \mu$ and variance $= \sigma^2$

THEN as $n \rightarrow \infty$

the sampling distribution of $\bar{X}_n = \left[\frac{\sum_{i=1}^n X_i}{n} \right]$ is eventually

Normal with mean $= \mu$ and variance $= \sigma^2/n$

In words:

“In the long run, averages have distributions that are well approximated by the Normal”

“The sampling distribution of \bar{X}_n , upon repeated sampling, is eventually Normal $\left(\mu, \frac{\sigma^2}{n} \right)$ ”

Source: PubHlth 540 Lecture Notes 5 (Normal Distribution), page 25

Nature

Population/
Sample

Observation/
Data

Relationships/
Modeling

Analysis/
Synthesis

We learned that, often (too often, sometimes), research focuses on the behavior of the sample mean

- Average length of hospital stay
- Average tumor size
- Average frequency of drug injections per week

According to the central limit theorem, the variance of the statistic \bar{X}_n (over its sampling distribution) is reasonably assumed to be:

$$\text{Variance}(\bar{X}_n) = \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$$

However, since we typically do not know the value of σ^2 . As our guess, we will use the following as our estimate:

$$\text{Var}(\hat{\bar{X}}_n) = \frac{S^2}{n}$$

where S^2 is the sample variance:

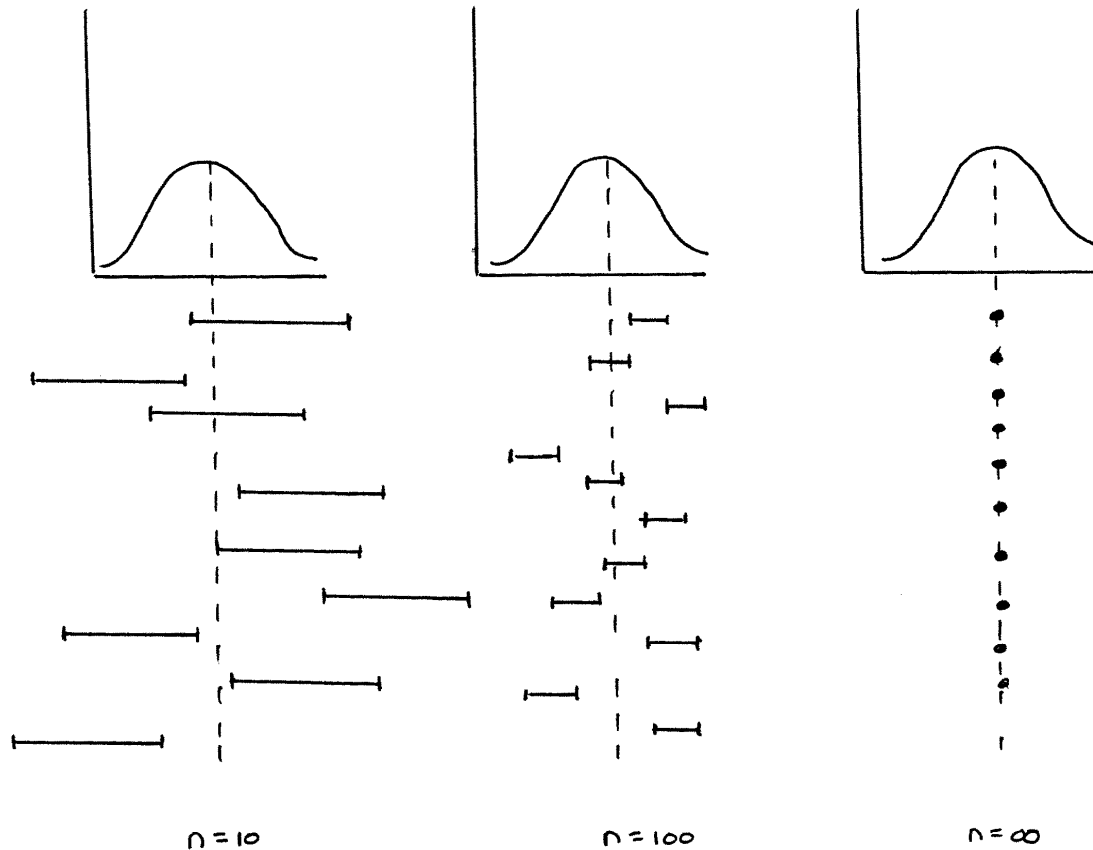
$$S^2 = \frac{\sum (X - \bar{X})^2}{n-1}$$

5. The Statistical Confidence Interval

“A confidence interval shows the likely range in which the [sample] mean would fall if the sampling exercise were to be repeated” (source: Crawley MJ. *Statistics: An Introduction Using R*. Wiley, 2005, page 45)

In *PubHlth 540 - Introductory Biostatistics*, we learned that a confidence interval expresses both the uncertainty with which we can estimate a population parameter (CI width) and the confidence we attach to the estimate (CI level) As the term implies, random sampling yields samples that are different from one to the next. As result, a single estimate of a population parameter will also differ from one sample to the next.

The following schematic might represent all possible 95% CI estimates of the mean, in 3 scenarios: (1) simple random sampling of sample sizes of $n=10$; (2) simple random sampling of sample sizes of $n=100$; and (3) simple random sampling of the entire source population.



All other things being equal, sampling designs with larger sample sizes or smaller measurement variability or both produce narrower (more precise) confidence intervals.



We also learned that a confidence interval estimate incorporates: (1) point estimate; (2) standard error of the point estimate; and (3) specified confidence level.

(1) One Sample, Normal Distribution

Confidence Interval for μ (σ^2 known)

$$\text{lower limit} = \bar{X} - z_{(1-\alpha/2)100} (\sigma / \sqrt{n})$$

$$\text{upper limit} = \bar{X} + z_{(1-\alpha/2)100} (\sigma / \sqrt{n})$$

For review, see: PubHlth 540 Lecture Notes 6 (Estimation), pp 33-36

(2) One Sample, Normal Distribution

Confidence Interval for μ (σ^2 NOT known)

$$\text{lower limit} = \bar{X} - t_{DF; (1-\alpha/2)100} (s / \sqrt{n})$$

$$\text{upper limit} = \bar{X} + t_{DF; (1-\alpha/2)100} (s / \sqrt{n})$$

For review, see: PubHlth 540 Lecture Notes 6 (Estimation), pp 38-40

(3) One Sample, Normal Distribution

Confidence Interval for σ^2

$$\text{lower limit} = \frac{(n-1)S^2}{\chi_{1-\alpha/2}^2}$$

$$\text{upper limit} = \frac{(n-1)S^2}{\chi_{\alpha/2}^2}$$

For review, see: PubHlth 540 Lecture Notes 6 (Estimation), p 41-43

(4) One Sample Proportion, Binomial Distribution
Confidence Interval for π

$$\text{lower limit} = \bar{X} - (z_{1-\alpha/2})\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}$$

$$\text{upper limit} = \bar{X} + (z_{1-\alpha/2})\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}$$

For $n \leq 30$ or so, use the following formulae instead:

$$\text{lower limit} = \bar{X} - (z_{1-\alpha/2})\sqrt{\frac{0.5(0.5)}{n}}$$

$$\text{upper limit} = \bar{X} + (z_{1-\alpha/2})\sqrt{\frac{0.5(0.5)}{n}}$$

For review, see: PubHlth 540 Lecture Notes 6 (Estimation), pp 60-63

(5) Paired Data Sample, Normal Distribution
Confidence Interval for μ_d (σ_d^2 known)

$$\text{lower limit} = \bar{d} - (z_{1-\alpha/2})\left(\sigma_d/\sqrt{n}\right)$$

$$\text{upper limit} = \bar{d} + (z_{1-\alpha/2})\left(\sigma_d/\sqrt{n}\right)$$

For review, see: PubHlth 540 Lecture Notes 6 (Estimation), pp 44-47 except use z percentile

(6) Paired Data Sample, Normal Distribution
Confidence Interval for μ_d (σ_d^2 NOT known)

$$\text{lower limit} = \bar{d} - (z_{1-\alpha/2})\left(s_d/\sqrt{n}\right)$$

$$\text{upper limit} = \bar{d} + (z_{1-\alpha/2})\left(s_d/\sqrt{n}\right)$$

For review, see: PubHlth 540 Lecture Notes 6 (Estimation), pp 44-47

(7) Paired Data Sample, Normal Distribution
Confidence Interval for σ_d^2

$$\text{lower limit} = \frac{(n-1)S_d^2}{\chi_{1-\alpha/2}^2}$$

$$\text{upper limit} = \frac{(n-1)S_d^2}{\chi_{\alpha/2}^2}$$

For review, see: PubHlth 540 Lecture Notes 6 (Estimation), page 48

(8) Two Independent Samples, Normal Distribution
Confidence Interval for $[\mu_1 - \mu_2]$

CI = [point estimate] ± (conf.coeff)SE[point estimate]			
Scenario →	σ_1^2 and σ_2^2 are both known	σ_1^2 and σ_2^2 are both NOT known but are assumed EQUAL	σ_1^2 and σ_2^2 are both NOT known and NOT Equal
Estimate	$\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}$	$\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}$	$\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}$
SE to use	$SE[\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}] = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$S\hat{E}[\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}] = \sqrt{\frac{S_{\text{pool}}^2}{n_1} + \frac{S_{\text{pool}}^2}{n_2}}$ where you already have obtained: $S_{\text{pool}}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$	$S\hat{E}[\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}] = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$
Conf Coeff: Use Percentiles from	Normal	Student's t	Student's t
Degrees freedom	Not applicable	$(n_1 - 1) + (n_2 - 1)$	$f = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left(\frac{\left[\frac{S_1^2}{n_1}\right]^2}{n_1 - 1} + \frac{\left[\frac{S_2^2}{n_2}\right]^2}{n_2 - 1}\right)}$

Source: PubHlth 540 Notes 6 (Estimation) page 53

(9) Two Independent Samples, Normal Distribution
Confidence Interval for σ_1^2/σ_2^2

$$\text{lower limit} = \left(\frac{1}{F_{n_1-1; n_2-1; (1-\alpha/2)}} \right) \frac{S_1^2}{S_2^2}$$

$$\text{upper limit} = \left(\frac{1}{F_{n_1-1; n_2-1; (\alpha/2)}} \right) \frac{S_1^2}{S_2^2}$$

For review, see: PubHlth 540 Lecture Notes 6 (Estimation), pp 57-59

(10) Two Independent Proportions, Binomial Distributions
Confidence Interval for $\pi_1 - \pi_2$

$$\text{lower limit} = (\bar{X} - \bar{Y}) - (z_{1-\alpha/2}) \sqrt{\frac{\bar{X}(1-\bar{X})}{n_x} + \frac{\bar{Y}(1-\bar{Y})}{n_y}}$$

$$\text{upper limit} = (\bar{X} - \bar{Y}) + (z_{1-\alpha/2}) \sqrt{\frac{\bar{X}(1-\bar{X})}{n_x} + \frac{\bar{Y}(1-\bar{Y})}{n_y}}$$

For $n_x \leq 30$ or $n_y \leq 30$ or so, use the following formulae instead:

$$\text{lower limit} = (\bar{X} - \bar{Y}) - (z_{1-\alpha/2}) \sqrt{\frac{0.5(0.5)}{n_x} + \frac{0.5(0.5)}{n_y}}$$

$$\text{upper limit} = (\bar{X} - \bar{Y}) + (z_{1-\alpha/2}) \sqrt{\frac{0.5(0.5)}{n_x} + \frac{0.5(0.5)}{n_y}}$$

For review, see: PubHlth 540 Lecture Notes 6 (Estimation), pp 64-66

6. Statistical Hypothesis Testing

In *PubHlth 540 - Introductory Biostatistics*, we learned that in statistical hypothesis testing, a null hypothesis model is applied to a set of data. Importantly, the data are sacrosanct. So, if the application of the null hypothesis to the given data leads to an unlikely conclusion (small p-value), this suggests rejection of the null hypothesis. “Of course, saying ‘we do not reject the null hypothesis’ and ‘the null hypothesis is true’ are two quite different things. For instance, we may have failed to reject a false null hypothesis because our sample size was too low, or because our measurement error was too large. Thus, p values are interesting, but they don’t tell the whole story; effect sizes and sample sizes are equally important in drawing conclusions.” (source: Crawley MJ. *Statistics: An Introduction Using R*. Wiley, 2005, page 4)

We learned that statistical hypothesis testing examines the distance (in standard error units!) between the observed and the null hypothesis expectation.

	<p>Scenario 1 – NULL Hypothesis is true $\Pr [\bar{X} \text{ is as extreme or more extreme than value observed}] = \text{large}$</p> <ul style="list-style-type: none"> • Observed sample mean <i>is close</i> to null mean. • Assuming the null hypothesis model leads to a relatively large probability that \bar{X} is the value actually observed or more extreme. (Note – extreme is always in the direction of the alternative) • Test statistic value = small P-value = large number → “do NOT reject the null” <p>Scenario 1 – NULL Hypothesis is false. Some distant ALTERNATIVE is true $\Pr [\bar{X} \text{ is as extreme or more extreme than value observed}] = \text{small}$</p> <ul style="list-style-type: none"> • Observed sample mean <i>is distant</i> from the null mean. • Assuming the null hypothesis model leads to a relatively small probability that \bar{X} is the value actually observed or more extreme. (Note – extreme is always in the direction of the alternative) • Test statistic value = large P-value = small number → “REJECT the null”
--	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Example of Statistical Hypothesis Test – Significance Level Method

Source: PubHlth 540 Unit 7 Notes (Hypothesis Testing) pp 12-15

Suppose that, with standard care, cancer patients are expected to survive a mean duration of time equal to 38.3 months. Investigators are hopeful that a new therapy will improve survival. Next, suppose that the new therapy is administered to 100 cancer patients. It is observed that they experience instead an average survival time of 46.9 months. Is the observed survival under the new treatment statistically significantly improved (relative to standard care)?

Identify the research question

With standard care, the expected survival time is $\mu = 38.3$ months. With the new therapy, the observed 100 survival times, X_1, X_2, \dots, X_{100} have average $\bar{X}_{n=100} = 46.9$ months. *Is this compelling evidence that $\mu_{\text{true}} > 38.3$?*

Assume the null hypothesis is true and state the corresponding null hypothesis probability model

For now, we'll assume that the 100 survival times follow a distribution that is Normal (Gaussian). We'll suppose further that it is known that $\sigma^2 = 43.3^2$ months².

Specify the null and alternative hypotheses

$$H_0: \mu_{\text{true}} = \mu_0 \leq 38.3 \text{ months}$$

$$H_A: \mu_{\text{true}} = \mu_A > 38.3 \text{ months}$$

Reason “proof by contradiction”

IF: the null hypothesis is true, so that $\mu_{\text{true}} = \mu_0 = 38.3$

THEN: what are the chances that a mean of 100 survival times will be “as extreme or more extreme than the value observed, namely 46.9?”

Nature

Population/
Sample

Observation/
Data

Relationships/
Modeling

Analysis/
Synthesis

Specify a “proof by contradiction” rule.

Statistically, assuming the null hypothesis in light of the observed data leads to an unlikely conclusion (translation: small p-value) if there is at most a small chance that the mean of 100 survival times is 46.9 or greater when its expected value is 38.3. We calculate the value of such chances as

$$\Pr[\bar{X}_{n=100} \geq 46.9 \mid \mu_{true} = \mu_o = 38.3]$$

P-value calculation of such chances presuming H_0 true.

Under the assumption that the null hypothesis is true:

X_1, X_2, \dots, X_{100} is a simple random sample from a Normal($\mu = 38.3, \sigma^2 = 43.3^2$).

This, in turn, says that under the assumption that the null hypothesis is true:

$\bar{X}_{n=100}$ is distributed Normal ($\mu = 38.3, \sigma^2 = 43.3^2 / (n = 100)$)

Recall ... How extreme is “extreme” is an example of “signal-to-noise”.

<p><u>Signal -</u> “46.9 is 8.6 months away from 38.3” Signal = 8.6</p> <p>Is 8.6 extreme or not?</p>	$(46.9 - 38.3) = 8.6$
<p>Noise - Noise is the scatter/variability of the average. We measure this using the SE</p> <p>How “noisy” is the mean typically? This is SE?</p>	$SE(\bar{X}_{n=100}) = \frac{\sigma}{\sqrt{100}} = \frac{43.3}{10} = 4.33$
<p><u>Signal-to-Noise (Z-score)</u> <u>Signal, in units of months, has been re-expressed in units of noise (SE units)</u></p> <p>“46.9 is 1.99 SE units away from 38.3”</p>	$\begin{aligned} Z\text{-score} &= \frac{(\bar{X}_{n=100} - \mu_{\bar{X} \text{ under NULL}})}{SE(\bar{X}_{n=100})} \\ &= \frac{(46.9 - 38.3)}{SE(\bar{X}_{n=100})} \\ &= \frac{8.6 \text{ months}}{4.33 \text{ months}} \\ &= 1.99 \text{ SE units} \end{aligned}$

Nature

Population/
Sample

Observation/
Data

Relationships/
Modeling

Analysis/
Synthesis

Z-score=1.99 says:

*“The observed mean of 46.9 is 1.99 SE units away from the **null hypothesis** expected value of 38.3”*

Logic of Proof-by-Contradiction says:

*“Under the assumption that the **null hypothesis is true**, there are 2 in 100 chances of obtaining a mean as far away from 38.3 as the value of 46.9”*

$$\Pr[\bar{X}_{n=100} \geq 46.9 | \mu_{true} = \mu_{null} = 38.3]$$

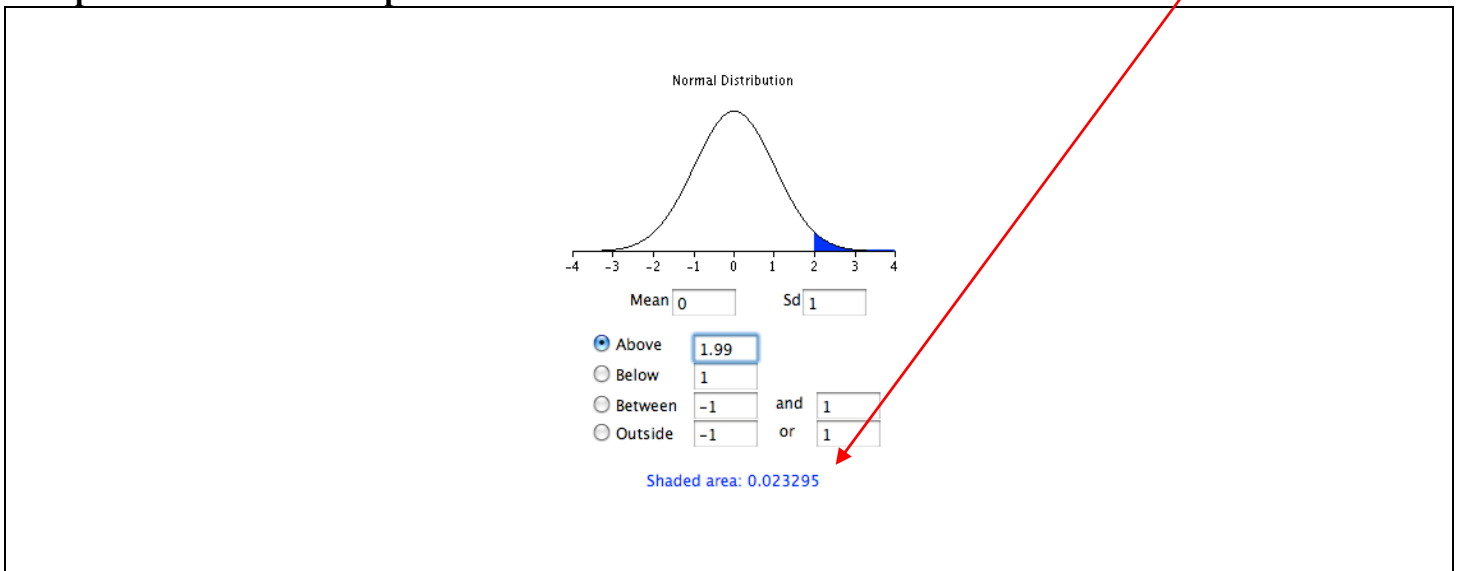
$$= \Pr[Z\text{-score} \geq 1.99] = .02$$

Statistical Reasoning of “likely” says:

*“If the null hypothesis, when examined in light of the data, leads us to something that is ‘unlikely’, namely a small p-value (**shaded area in blue below**), then the null hypothesis is severely challenged, if not contradicted. →*

Statistical rejection of the null hypothesis.

Graphical illustration of a p-value calculation



Example of Statistical Hypothesis Test – Critical Region Approach

Source: PubHlth 540 Unit 7 Notes (Hypothesis Testing) pp 31-34

- We agree *in advance (prior to collecting data)* that we will judge “extreme” values as inconsistent with the null hypothesis and then reject the null hypothesis, even though null hypothesis extreme values are theoretically possible.
- Should such “extreme” occur and we *incorrectly reject a true null hypothesis* we will have made a *type I error*.
- In developing a critical region test, we determine, ahead of time, the set of extreme values (this is called the *critical region*) that will prompt (rightly or wrongly!) rejection of the null hypothesis.

With standard care, cancer patients are expected to survive a mean duration of time equal to 38.3 months. Hypothesized is that a new therapy will improve survival. In this study, the new therapy is administered to 100 cancer patients. Their average survival time is 46.9 months. Suppose σ^2 known = 43.3^2 months squared. Is this statistically significant evidence of improved survival *at the 0.05 level*?

Null Hypothesis Probability Model Assumptions.

X_1, X_2, \dots, X_{100} is a simple random sample from a Normal($\mu, \sigma^2 = 43.3^2$)

Null and alternative hypotheses

$H_0: \mu_{true} = \mu_0 \leq 38.3$ months

$H_A: \mu_{true} = \mu_A > 38.3$ months

The appropriate Test Statistic is a Z-Score

The null hypothesis gives us the following:

- X_1, X_2, \dots, X_{100} is a simple random sample from a Normal($\mu = 38.3, \sigma^2 = 43.3^2$).
- $\bar{X}_{n=100}$ is distributed Normal ($\mu = 38.3, \sigma^2 = 43.3^2/100$)
- Again, we'll use as our test statistic the z-score standardization of $\bar{X}_{n=100}$, obtained under the assumption that the null hypothesis is correct.

$$\text{Test Statistic} = \text{z-score} = \frac{\bar{X}_{n=100} - \mu_{null}}{SE(\bar{X}_{n=100})}$$

Nature

Population/
Sample

Observation/
Data

Relationships/
Modeling

Analysis/
Synthesis

Using the direction of the alternative, obtain the 0.05 critical region

Step 1: Identify what is meant by “extreme” or “critical”:

In this example, the alternative is **one sided** and extreme values in the direction of the alternative are **large positive** values of the pivotal quantity.

Step 2: Solve for the critical region of the pivotal quantity:

In this example, solve for the range of extreme values of a Z-score random variable distributed Normal(0,1) such that the area under the null hypothesis curve in the direction of large positive (right tail) is 0.05.

Source: http://davidmlane.com/hyperstat/z_table.html

Be sure to scroll down to the second calculator that is provided.

<p>Normal Distribution</p> <p>Mean: <input type="text" value="0"/> Sd: <input type="text" value="1"/></p> <p>Shaded Area: <input type="text" value=".05"/></p> <p><input checked="" type="radio"/> Above: 1.6449 <input type="radio"/> Below <input type="radio"/> Between <input type="radio"/> Outside</p>	<p>Step 1: Enter .05 for shaded area</p> <p>Step 2: Select radio button “Above”</p> <p>Step 3: Read critical region as z-score ≥ 1.6449</p>
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Step 3: Solve for the critical region of \bar{X} :

$$\text{Pivotal Quantity} = \text{z-score} = \frac{\bar{X}_{n=100} - \mu_{\text{null}}}{\text{SE}(\bar{X}_{n=100})} \geq 1.6449 \rightarrow$$

$$\frac{\bar{X}_{n=100} - 38.3}{4.33} \geq 1.6449 \rightarrow$$

The critical region is $\bar{X}_{n=100} \geq 45.42$

Step 4: Interpret:

In words, “this one sided .05 test of the null versus alternative hypotheses rejects the null hypothesis for critical region values of $\bar{X}_{n=100} \geq 45.42$.

Examine the observed to see if it is in the critical region

$\bar{X}_{n=100} = 46.9$ is in the critical region because it is greater than 45.42.

Interpret.

Because $\bar{X}_{n=100} = 46.9$ and is in the critical region, it is significant at the 0.05 level. According to the critical region approach with type I error = 0.05, **reject the null hypothesis.**

We learned that statistical significance is NOT biological significance***Question:***

It is observed that better results are obtained for patients receiving treatment “A” than treatment “B.” What are the possible explanations?

Answer:

- Treatment “A” is truly superior; or.
- Groups “A” and “B” were not comparable initially; or
- Treatment “B” is actually superior so that the data reflect instead an event of low probability.

... and we learned how to do some one and two sample hypothesis tests.

(1) One Sample, Normal Distribution

Test of μ (σ^2 known)

Z-test

For review, see: PubHlth 540 Lecture Notes 7 (Hypothesis Testing), pp 31-34

(2) One Sample, Normal Distribution

Test of μ (σ^2 NOT known)

Student t-test

For review, see: PubHlth 540 Lecture Notes 7 (Hypothesis Testing), pp 35-37

(3) One Sample, Normal Distribution

Test of σ^2

Chi Square test

For review, see: PubHlth 540 Lecture Notes 7 (Hypothesis Testing), pp 38-40

(4) One Sample Proportion, Binomial Distribution

Test of π

Z-test

For review, see: PubHlth 540 Lecture Notes 7 (Hypothesis Testing), pp 53-56

(5) Paired Data Sample, Normal Distribution

Test of μ_d (σ_d^2 known)

Z-test

For review, see: PubHlth 540 Lecture Notes 7 (Hypothesis Testing), pp 41-43

(6) Paired Data Sample, Normal Distribution**Test of μ_d (σ_d^2 NOT known)****Student t-test**

For review, see: *PubHlth 540 Lecture Notes 7 (Hypothesis Testing)*, page 44

(7) Paired Data Sample, Normal Distribution**Test of σ_d^2** **Chi Square Test**

For review, see: *PubHlth 540 Lecture Notes 7 (Hypothesis Testing)*, pp 38-40.

(8) Two Independent Samples, Normal Distribution**Test of $|\mu_1 - \mu_2|$**

Z-test or Student t-test, depending on what's known about the variances.

For review, see: *PubHlth 540 Lecture Notes 7 (Hypothesis Testing)*, pp 45-49

(9) Two Independent Samples, Normal Distribution**Test of Equality of σ_1^2 and σ_2^2** **F-test**

For review, see: *PubHlth 540 Lecture Notes 7 (Hypothesis Testing)*, pp 50-52

(10) Two Independent Proportions, Binomial Distributions**Confidence Interval for $\pi_1 - \pi_2$** **Z-test**

For review, see: *PubHlth 540 Lecture Notes 7 (Hypothesis Testing)*, pp 57-59