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 the same 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 (whatever that is!). 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 tools for assessing chance, effect size, and meaningful inference.

Your introductory biostatistics or statistics course (BIOSTATS 540?) introduced 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 vales are separated by gaps, the latter meaning that, between two values (eg "1 visit" versus "2" visits", nothing in-between is possible);

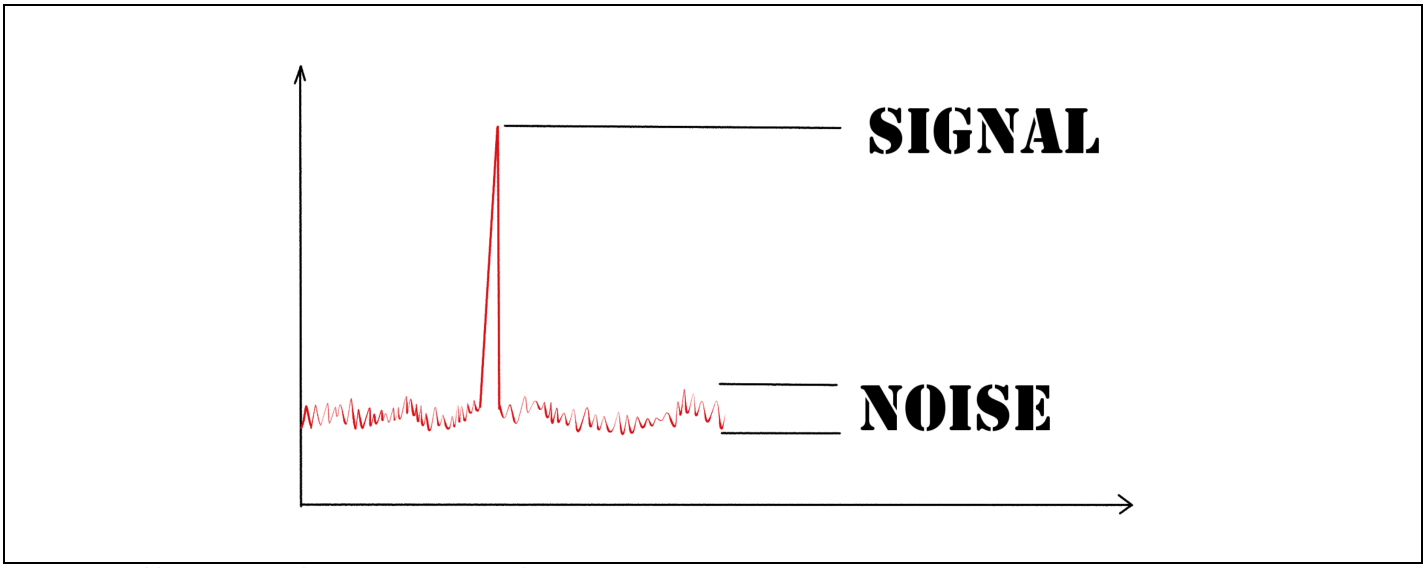
(3) For a 95% ***confidence interval (CI)***, we are 95% confident that our random sampling captured the unknown true population parameter value (which is not random); and

(4) Assessing the likelihood of a particular result (or a result more extreme), under a presumed null hypothesis (***p-value***), is done using a probability model.

Table of Contents

1. Signal to Noise

Variation is everywhere. Two observations might be different because measurement itself is "inherently variable" and this can be for a variety of reasons (**noise**). Or, the discrepancy between two observations some real, "systematic", difference (**signal**). So which is it? Signal or noise?



Source: https://www.fragile.at/your-signal-noise-ratio/

Noise	Signal
Example – Blood pressure. **Yesterday: 110/65 mm Hg.** **Today: 105/70 mm Hg** ⟷	**Example – Blood pressure.** **Yesterday: 110/65 mm Hg.** **Today: 180/120 mm Hg**
Not so different. The discrepancy is small. I will interpret this as reflecting *"noise"*, an instance of *natural fluctuations* one might expect from day to day and I don't obsess about its cause (e.g., today I had an extra cup of coffee).	Very different! The discrepancy is large. I will interpret this as evidence of a *"signal"*, an instance of something systematic that has produced the discrepancy. (e.g., I have developed clinical hypertension)

It is often the "systematic" variations that interest us. A "systematic" variation, if it is big enough, becomes interesting to us as a "signal", a discovery of something systematic. So, how big is big enough"? We compare "signal" to "noise" by looking at its ratio = [signal] / [noise].

We need a tool for distinguishing "systematic/signal" versus "noise"; one tool is the signal-to-noise ratio. In BIOSTATS 540, we learned that a statistical test is a measure of [signal] / [noise]. We learned that the anatomy of some statistics is exactly a signal-to-noise ratio:

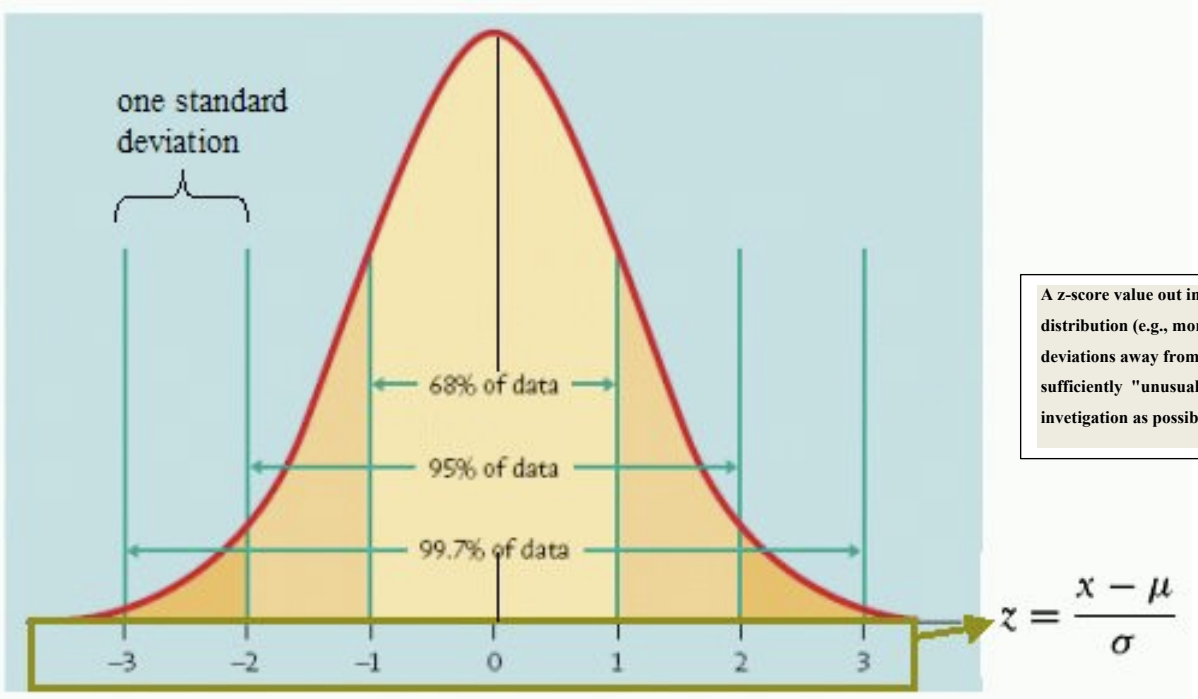
- **Signal** = [Observed e.g., average response, $\overline{X}$] - [Null hypothesis expected; e.g.μ_{NULL}]
- **Noise** = Standard Error of the observed; e.g., $SE\ (\overline{X})$

Many statistical tests, upon closer look, are easily seen to be signal-to-noise statistics.

Two examples are the Z-test and T-test, also called Z-Score and T-Score = $\frac{\text{(observed - expected|null)}}{\text{se(observed)}} = \frac{\text{signal}}{\text{noise}}$

Nifty! A Z-score (or a T-score) is a measure of signal to noise (a "metric" of unusualness). What's more (and this is really great), often, Z-scores are distributed standard normal. So we can use the Normal distribution curve as our reference, against which we can gauge whether somethine is unusual enough to call it a "signal" versus just "noise.

Plot of the Standard Normal Distribution
(source: Google images)



A z-score value out in the tail of this distribution (e.g., more than 3 standard deviations away from 0) will be declared sufficiently "unusual" to warrant further invetigation as possibly a signal.

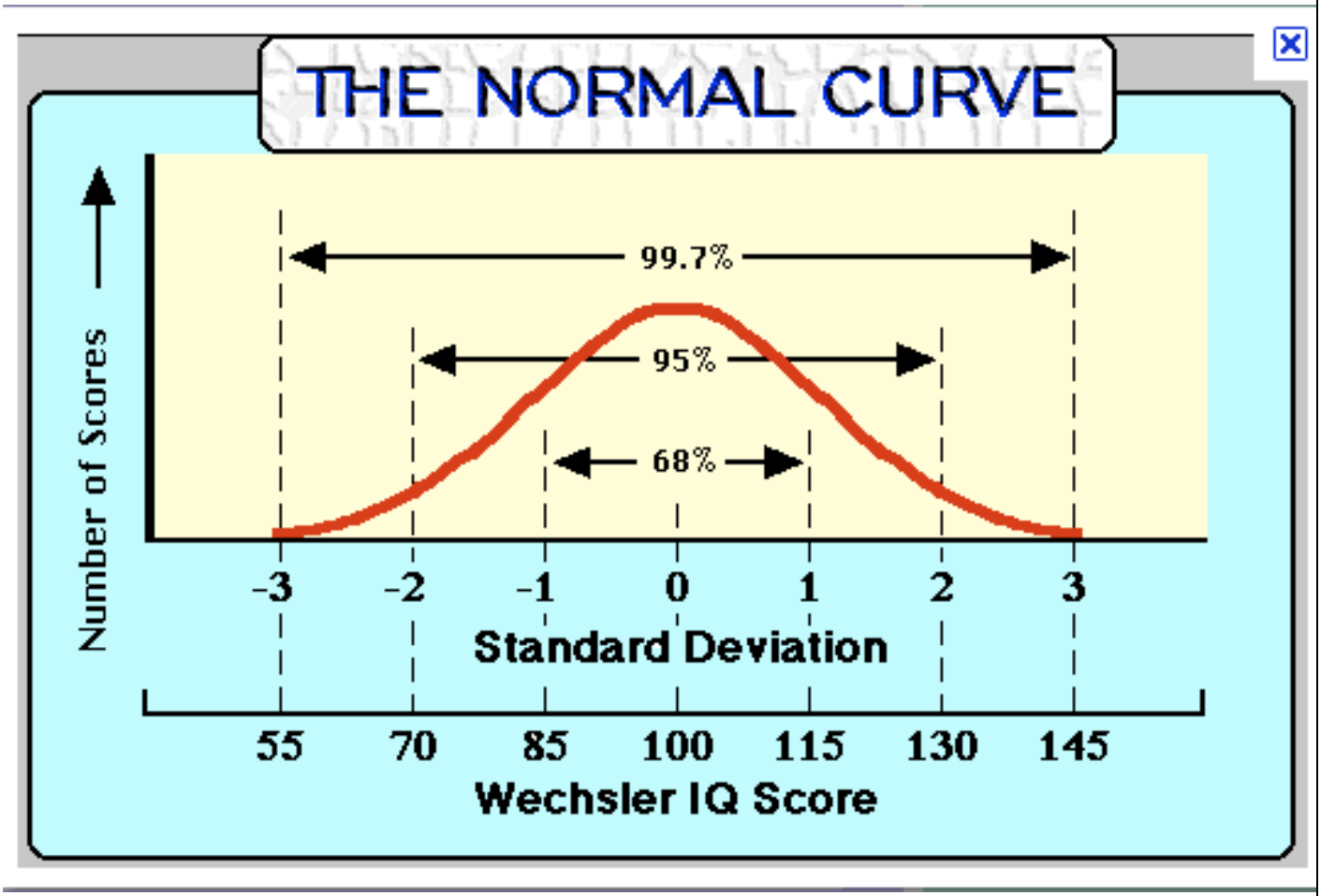
The standard normal (Gaussian) distribution is centered at its mean value μ=0.
Along the horizontal axis are tick marks at increments of 1 standard deviation (σ=1)

*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. We might intepret these as **"signals"** that warrant further investigation, even though, theoretically, such values are possible.*

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 standard deviation 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, while theoretically possible in the general population, are unlikely as they occur in 5% or less of the population. So, when we actually encounter such extremes, we're inclined to consider the possibility that these values are actually from some **other** population; such these values are then regarded as "signals" warranting further study.*

"How big is big enough" is not just about size. It's also about whether or not it matters.
"Signal-to-noise" is a matter of perspective. "Noise" to one person might be "signal" to another.

Example.

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.

- ***What matters to you ….***
  From a ***public health perspective***, the variability in outcomes in the experimental versus control groups reflect noise; whereas,

- ***may be different than what matters to me***
  For the ***individual patient***, any extra life afforded by the new treatment is hugely significant (signal).

2. Description and Estimation

Description. In *BIOSTATS 540, Introductory Biostatistics*, we learned how to construct 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			
	Qualitative		Quantitative	
Type	Nominal	Ordinal	Discrete	Continuous
Example	**religion**	**strength of opinion**	**# visits to doctor**	**weight**
Visualization	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 Relative Frequency Cumulative Frequency	Frequency Relative Frequency Cumulative Frequency means, variances, percentiles	- - - means, variances, percentiles

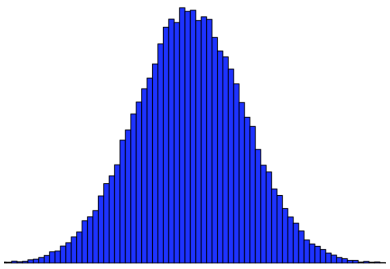
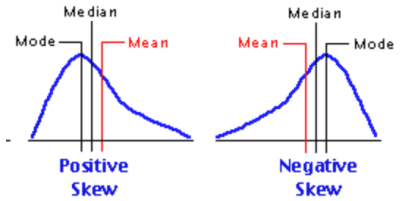
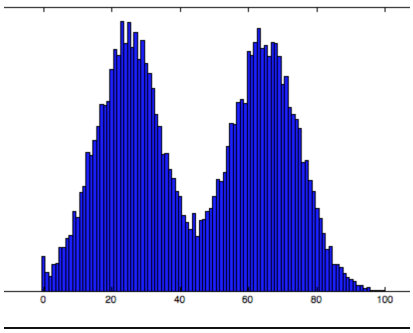
Source: BIOSTATS 540 Lecture Notes 1 (Summarizing Data), page 12.

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*)

There is also more than one choice for describing dispersion

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

		
Symmetric	**Skewed**	**Bimodal**

Think about it ….

For each of these 3 scenarios ("Symmetric", "Skewed", "Bimodal"):

1. Is the sample mean a good choice for summarizing central location?
2. Would the sample median be a better choice?
3. And, actually, does either make sense?

We learned the distinction between sample variance (S^2) and sample standard deviation (SD) .

Sample variance (S^2) S^2 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 instead of by n) distance from the sample mean.

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

Standard Deviation (S or SD). S (or SD) is the square root of S^2. The advantage of the square root operation is that the resulting summary has the same scale as the original values.

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

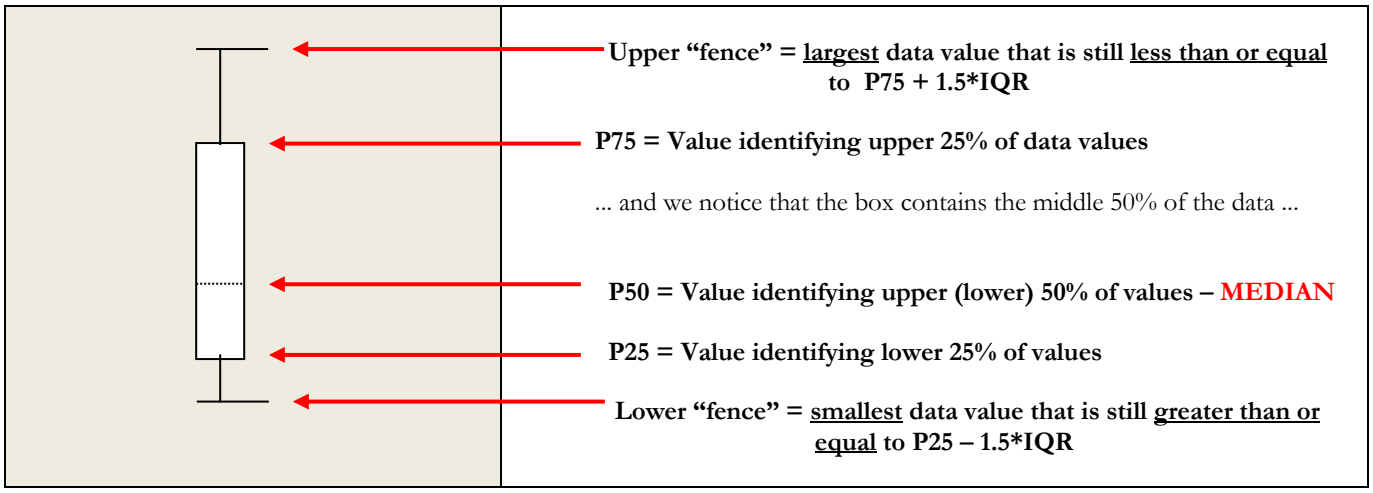
Illustration of Calculation

Source: BIOSTATS 540 Lecture Notes 1 (Summarizing Data), page 28.

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**

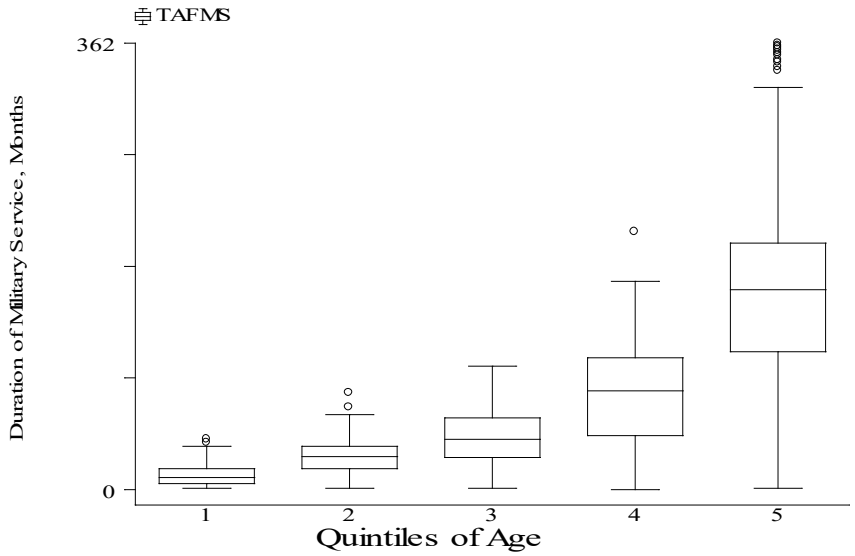
- $\sum_{i=1}^{11} X_i = 1771$ **days** → **Mean for sample is** $\bar{X} = \frac{1771}{11} = 161$ **days**
- **Sample variance is** $s^2 = \frac{\sum_{i=1}^{11}\left(X_i - \bar{X}\right)^2}{n-1} = \frac{285236}{10} = 28523.6$ **days squared**
- **Sample standard deviation is** $s = \sqrt{s^2} = \sqrt{28523.6} = 168.89$ **days**

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



- The central box has P_{25} and P_{75} for its limits. It spans the middle half of the data.
- Recall. The IQR = "interquartile range" = $(P_{75} - P_{25})$ and is the size of the span of the "middle" 50%
- The line within the box identifies the median, P_{50}. 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". Asterisks (or dots) appearing beyond the fences (you may or may not have these) denote extreme values.
- Upper "fence" = largest observed data value that is $\leq$ $P_{75} + 1.5*IQR$
  = largest observed data value that is $\leq$ $P_{75} + 1.5*(P_{75} - P_{25})$.
- Lower "fence" = smallest observed data value that is $\geq$ $P_{25} - 1.5*IQR$
  = smallest observed data value that is $\geq$ $P_{25} - 1.5*(P_{75} - P_{25})$.

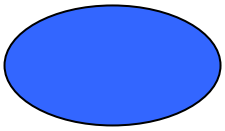
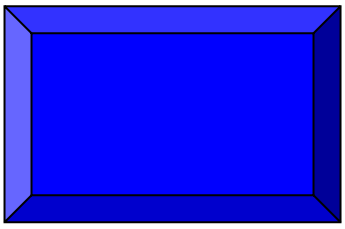
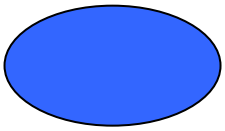
Example. Duration of Military Service by Age



- Duration of military service increases with age (key - the location of the medians are increasing with age).
- With age, the variability in duration of military service is greater. (key - the size of the boxes increase)
- The individual circles represent extreme values.

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

When the sample is representative of the population from which it was drawn, the values of the summary statistics computed from the sample are meaningful estimates of the corresponding population parameter values.

Description of a Sample	Estimation
sample 	**population**   **simple random sample "representative"** **sample** 
In description, there is no larger context than the available data.	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.

3. Just so it's clear – SD versus SE

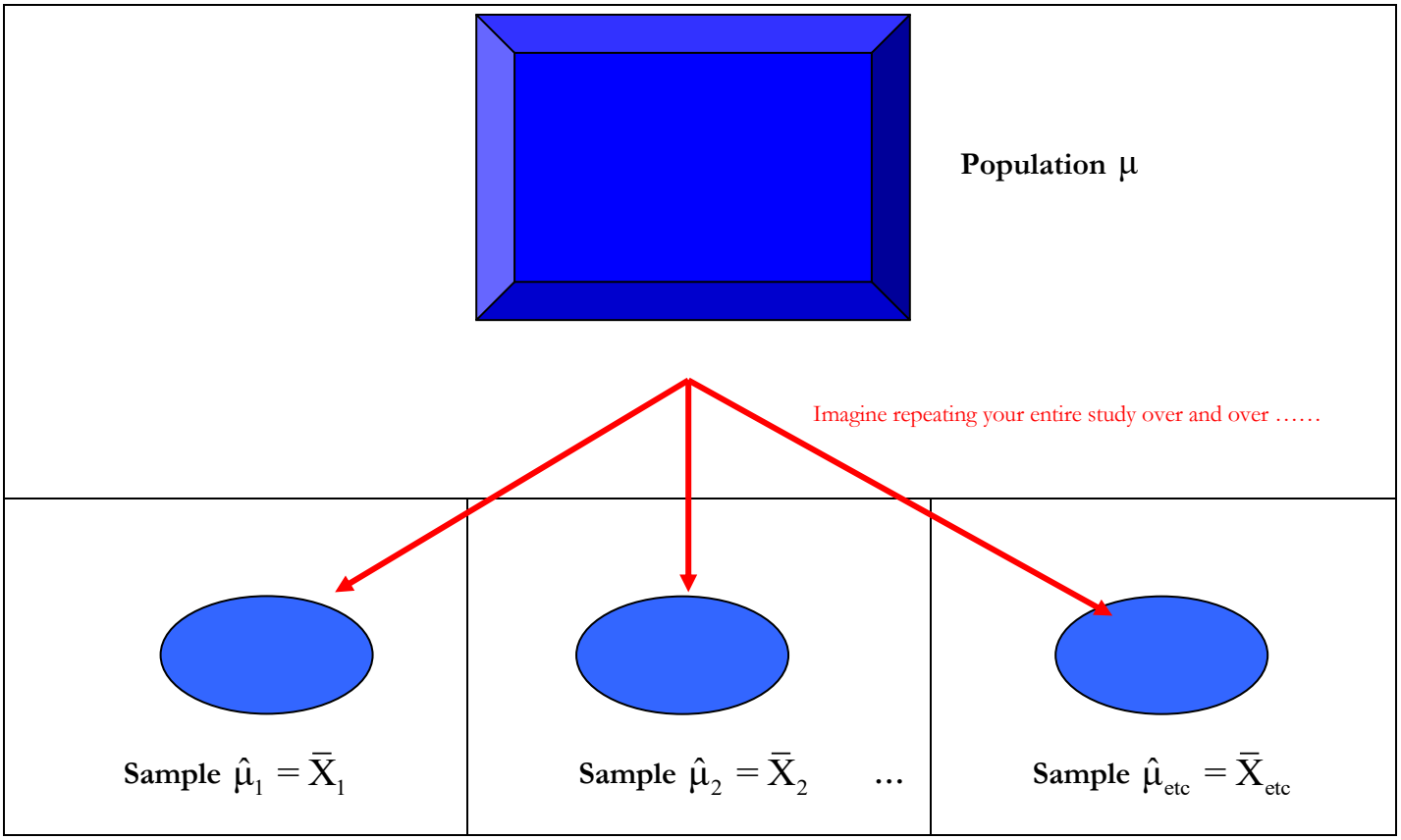
- **The standard deviation (SD or S) is a summary of the variability of <u>individuals in nature;</u> whereas**
- **the standard error (SE) is a summary of the variability of a <u>summary statistic among many replications of your study</u>** *Recall. We imagine a collection of values of a sample statistic such as the sample mean that is obtained by replicating your whole study over and over again. Spoiler – this collection will be called a sampling distribution.*

One SE (statistics) is the SE(mean), We clarified standard deviation SD versus a particular standard error SE, namely: the SE <u>of the mean $\bar{X}$</u>

<u>SD or S: Standard Deviation</u>	**<u>SE($\bar{X}$): Standard Error of the Mean</u> $\bar{X}$**
• Standard deviation is a measure of the variation of values (eg – cholesterol) from *individual* to *individual.*	• The standard error is a measure of the variation of the values of a statistics (eg – sample mean $\bar{X}$) from *sample* to *sample* (e.g., the sampling distribution of the sample mean $\bar{X}$).
• The standard deviation of the values of a (eg – cholesterol) in a ***population*** is represented by the symbol σ	• 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($\bar{X}$).
• The standard deviation among the values of a measurement (eg – cholesterol) in a ***sample*** is represented by either or two symbols: $\hat{\sigma}$ or S	***Example -*** Standard error of $\bar{X}$ • Under simple random sampling from a population with mean μ and variance σ^2, • If σ^2 is KNOWN, then use $SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ where n= size of sample. • But if σ^2 is NOT known, then use the estimate $\hat{SE}(\bar{X}) = \frac{s}{\sqrt{n}}$ where S = sample SD Recall. The little hat "^" (it's called a caret) tells us that this is an estimate.

We learned the meaning of a sampling distribution.

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 all these infinitely many $\bar{X}$ is analogous to a "population" but we don't call it a population. Instead, we call it the **sampling distribution** of $\bar{X}$.



Notes -

1). This picture is a schematic of the sampling distribution of $\bar{X}$.

2). The same approach could be used to generate the sampling distribution of any statistic we like; e.g.,

- sample median
- sample variance
- an estimated slope from a regression
- and so on and so on.

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.

- The pattern of occurrence of <u>many phenomena in nature</u> is described well using a normal distribution model.

- Even when the phenomena in a sample distribution are not described well by the normal distribution, the collection of all possible $\bar{X}$, known as the sampling distribution of <u>sample averages</u> obtained by repeated sampling from the parent distribution is eventually correctly described using the normal distribution (*Central limit theory*).

Normal Distribution (μ, σ^2)

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

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

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 $[X - \mu]^2$

Recall! $\frac{x-\mu}{\sigma} = \frac{observed - expected}{standard\ deviation} = z - score$

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

A random variable Z modelled as 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 a simple random sample of n independent observations $X_1 \ldots X_n$; and
2) the $X_1 \ldots X_n$ are each a random draw from the same distribution, *whatever that is.*; and
3) this distribution has mean = μ and variance = σ^2

THEN as n $\to\infty$

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

Normal with mean = μ and variance = σ^2/n

In words:

"In the long run, the sampling distribution of averages have distributions that are well approximated by the Normal distribution"

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

Source: BIOSTATS 540 Lecture Notes 7 (Normal Distribution), page 32.

We learned that, often, we want to communicate to our reader the "average" of what we observed (the sample mean)

For example, the sample mean might be:

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

The central limit theorem is all about the "average"

According to the central limit theorem, the variance of the statistic $\bar{X}_n$ (over its sampling distribution) is eventually equal to the following, and in the meantime can reasonably be assumed to be equal to: $$\text{Variance}(\bar{X}_n) = \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$$	However, since we typically do not know the value of σ^2, we will have to use a guess of σ^2 which we will write as $\hat{\sigma}^2$ (using the caret, see * below) as our estimate: $$\hat{\text{Var}}(\bar{X}_n) = \frac{\hat{\sigma}^2}{n} = \frac{S^2}{n}$$ where S^2 is the sample variance: $$S^2 = \frac{\sum(X-\bar{X})^2}{n-1}$$ * Recall. The little hat "^" (it's called a caret) tells us that this is an estimate.

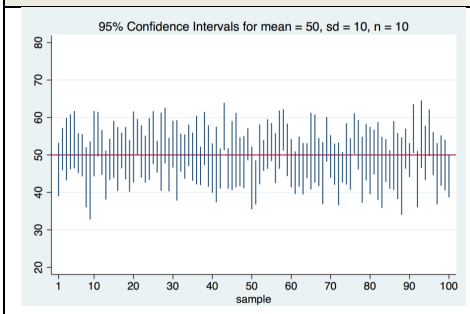
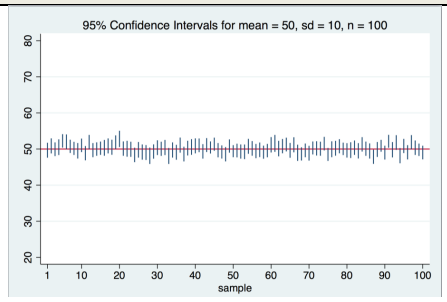
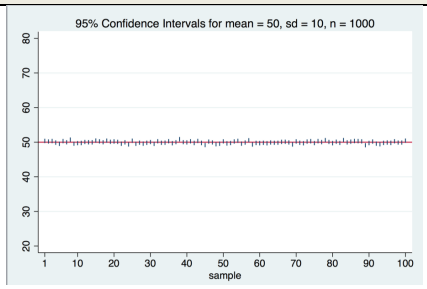
5. The Statistical Confidence Interval

"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 *BIOSTATS 540 - Introductory Biostatistics*, we learned that a confidence interval (CI) expresses both the uncertainty with which we can estimate a population parameter (CI width) and the confidence we attach to the confidence interval 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 can reasonably be expected to differ from one sample to the next. For example, $\bar{X}$ from one sample is not equal to $\bar{X}$ from the next.

The following schematic is taken from the BIOSTATS 540 lecture notes for Unit 8, Statistical Literacy. It represents 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 sample sizes n=1000. In all three scenarios, the sampling is from the same normal distribution having mean=50 (indicated by the red line) and standard deviation = 10.

n = 10	n = 100	n = 1000
		

Source: Fall 2020 BIOSTATS 540, Unit 8 – Statistical Literacy, page 47.

These pictures remind us that …

(1) Any one confidence interval either contains μ or it does not. Okay, this is hard to see, but perhaps you can? This illustrates that it is <u>incorrect</u> to say "*There is a 95% probability that the confidence interval contains μ*"

(2) For a given sample size (here, n=10), the width of all the confidence intervals is the same.

(3) As the sample size increases, the confidence intervals are more narrow (*more precise)*

(4) As n → infinity, μ is in the interval every time.

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)

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

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

For review, see: BIOSTATS 540 Lecture Notes 9 (One Sample Inference) pp 5-8.

(2) One Sample, Normal Distribution
Confidence Interval for μ (σ^2 NOT known)

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

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

For review, see: BIOSTATS 540 Lecture Notes 9 (One Sample Inference) pp 16-18.

(3) One Sample, Normal Distribution
Confidence Interval for σ^2

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

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

For review, see: BIOSTATS 540 Lecture Notes 9 (One Sample Inference) pp 21-23.

(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: BIOSTATS 540 Lecture Notes 9 (One Sample Inference) pp 47-50.

(5) Paired Data Sample, Normal Distribution
Confidence Interval for μ_d (σ^2_d 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: BIOSTATS 540 Lecture Notes 9 (One Sample Inference) pp 33-34

(6) Paired Data Sample, Normal Distribution
Confidence Interval for μ_d (σ^2_d 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: BIOSTATS 540 Lecture Notes 9 (One Sample Inference) pp 36-38.

(7) Paired Data Sample, Normal Distribution

Confidence Interval for σ^2_d

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

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

For review, see: BIOSTATS 540 Lecture Notes 9 (One Sample Inference) pp 44-46.

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

CI = [point estimate] ± (confidence coefficient)*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	$SE\left[\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}\right] = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$\hat{SE}\left[\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}\right] = \sqrt{\frac{S_{pool}^2}{n_1} + \frac{S_{pool}^2}{n_2}}$ **where you already have obtained:** $S_{pool}^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1)+(n_2-1)}$	$\hat{SE}\left[\bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}\right] = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$
Conf Coeff: Use Percentile from	**Normal**	**Student's t**	**Student's t**
Degrees freedom	**Not applicable**	$(n_1 - 1) + (n_2 - 1)$	$f = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left[\frac{S_1^2}{n_1}\right]^2}{n_1 - 1} + \frac{\left[\frac{S_2^2}{n_2}\right]^2}{n_2 - 1}}$

Source: BIOSTATS 540 Notes 10 (Two Sample Inference) pp 16-22.

(9) Two Independent Samples, Normal Distribution
Confidence Interval for σ^2_1/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: BIOSTATS 540 Lecture Notes 10 (Two Sample Inference) pp 27-30.

(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 review, see: BIOSTATS 540 Lecture Notes 10 (Two Sample Inference) pp 35-36.

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

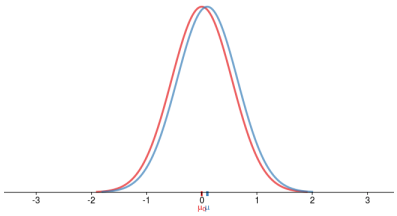
$$\text{lower limit} = (\bar{X} - \bar{Y}) - (z_{1-\alpha/2})\sqrt{\frac{0.5(0.5)}{n_x} + \frac{0.5(0.5)}{n_y}}$$

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

For review, see: BIOSTATS 540 Lecture Notes 10 (Two Sample Inference) pp 35.

6. Statistical Hypothesis Testing

In *BIOSTATS 540 - Introductory Biostatistics*, we learned that if assumption of, and application of, the provisional (null hypothesis) model to a set of data suggests that a very unlikely result has occurred (small p-value), this prompts us to reconsider the reasonableness of the provisional (null hypothesis) model (statistical 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)*



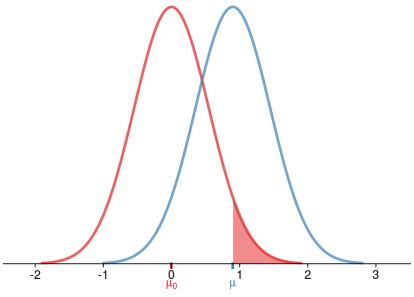
Step 1 –Begin by assuming that the null hypothesis is true

Under the null hypothesis assumption model, the two probability distributions ("true" and "null hypothesis") are identical or very nearly the same. This is why the two curves are right on top of each other. ***Note –*** We have not yet taken into consideration the given observed data. This comes next (middle picture).

$\overline{X}$

Step 2 – Consider the observed sample mean.

This picture represents the given summary statistic of the data. Notice that there is no probability distribution shown. ***Remember -*** we don't actually know which distribution gave rise to the data.



Step 3 – Argue "yes" or "no" consistency of the null hypothesis model assumption with the data.

In this picture, the **"true"** distribution that gave rise to the data is on the right. The **null hypothesis assumption model** is on the left. The shaded area is a probability calculation under the assumption that the null is true:

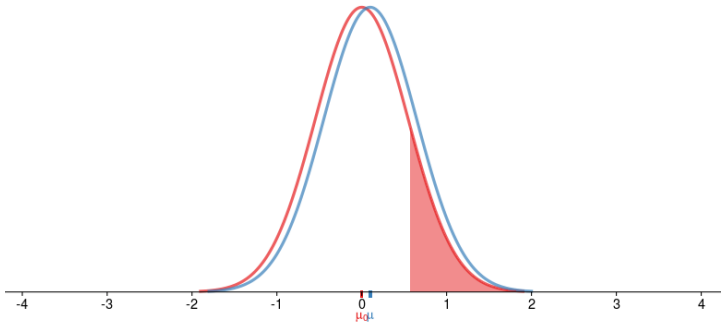
$$\Pr[\overline{X} \geq \text{observed} \mid \text{assuming null model}].$$

It answers the question "*Under the assumption of the null hypothesis, what are the chances of a value of the sample mean as extreme, or more, than was observed?*"

Small probability says "Assuming the null led to an unlikely event"
Large probability says "Assuming the null led to a likely event" $\overline{X}$

Source: https://istats.shinyapps.io/power/

A closer look at Pr [$\bar{X}$ ≥ observed value | assuming null hypothesis model is true]

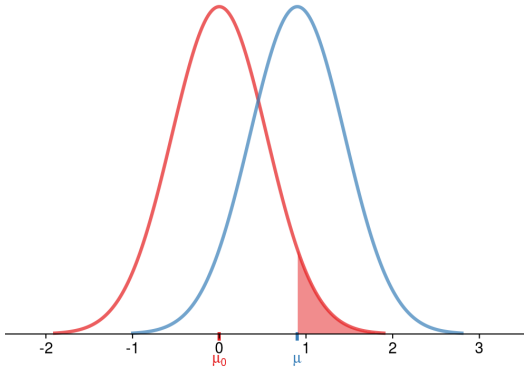


Scenario 1 - NULL is true

Pr [$\bar{X}$ ≥ observed value | assuming null is true]

= large

- The sample mean tends to be close to its "true" expected value, and this is also close to the null hypothesis model expected value. .
- The null hypothesis model probability that the sample mean $\bar{X}$ is as far away from the null hypothesis mean, or more extreme, is a large probability (large shaded area). That is - assuming the null hypothesis model leads to a likely outcome.
- Statistical decision - "do NOT reject the null".



Scenario 2 - ALTERNATIVE is true

Pr [$\bar{X}$ ≥ observed value | assuming null is true]

= small

- Observed sample mean is ***not*** close to null mean. Rather it tends to be close to its true mean which is now the alternative
- The null hypothesis model probability that the sample mean $\bar{X}$ is as far away from the null hypothesis mean, or more extreme, is a small probability (small shaded area). That is - assuming the null hypothesis model leads to an unlikely outcome.

  Statistical decision - "REJECT the null".

Source: https://istats.shinyapps.io/power/

Example of Statistical Hypothesis Test – Significance Level Method

Source: BIOSTATS 540 Unit 8 Notes (Statistical Literacy) pp 19-23.

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, $X_1, X_2, ..., X_{100}$ have average $\bar{X}_{n=100} = 46.9$ months. *Is this compelling evidence that* $\mu_{true} > 38.3$ *?*

Provisionally entertain that the null hypothesis is true and state the corresponding null hypothesis probability model. Then, later, use this model to obtain the null hypothesis "likelihood" of a particular result (the test statistic value) or one that is more extreme ("more unfavorable to the null").

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 $\sigma^2 = 43.3^2$ months squared.

Specify the null and alternative hypotheses

H_O: $\mu_{true} = \mu_O \leq 38.3$ **months**

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

Reason "proof by contradiction"

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

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?

Specify a "proof by contradiction" rule.

We will say that a "contradiction" of the null has occurred if assuming the null hypothesis to be true and applying it to the observed data leads to an unlikely conclusion (translation: small p-value). Put another way, we will say that a "contradiction" has occurred if there is at most a small null hypothesis chance that the mean of 100 survival times is 46.9 or greater when its (null hypothesis) expected value is 38.3. We calculate the value of such chances, and call this our p-value:

$$\text{p-value} = \Pr\left[\bar{X}_{n=100} \geq 46.9 \mid \mu_{true} = \mu_O = 38.3\right]$$

P-value calculation of such chances presuming H_O true.

Under the assumption that the null hypothesis is true:

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

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

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

Recall …How extreme is "extreme" is an example of "signal-to-noise" and we measure it in SE units.

Signal - **"46.9 is 8.6 months away from 38.3"** **Signal = 8.6 months** **Is 8.6 months noteworthy/extreme or not?**	$(46.9 - 38.3) = 8.6$ months
***Noise* –** **Noise is the scatter/variability of the average. We measure this using the SE units (instead of units of months)** **How "noisy" is the mean typically? We measure this noisiness (how far away it is from the null hypothesis mean) in units of SE.**	$1 \text{ SE}(\bar{X}_{n=100}) \text{ unit } = \frac{\sigma}{\sqrt{100}} = \frac{43.3}{10} = 4.33$ months
Signal-to-Noise (Z-score) ***Signal, in units of months, has been re-expressed in units of noise (SE units)*** **"46.9 is 1.99 SE units away from the (null hypothesis) expected survival of 38.3 months."**	$Z - score = \frac{\bar{X}_{n=100} - \mu_{\bar{X}_{n=100}}}{SE(\bar{X}_{n=100})} = \frac{(46.9 - 38.3)}{SE(\bar{X}_{n=100})} = \frac{8.6 \text{ months}}{4.33 \text{ months}/1 \text{ SE UNIT}} = 1.99 \text{ SE units}$

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 were 2 in 100 chances of obtaining a mean as far away from 38.3 months as the value of 46.9 months"

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

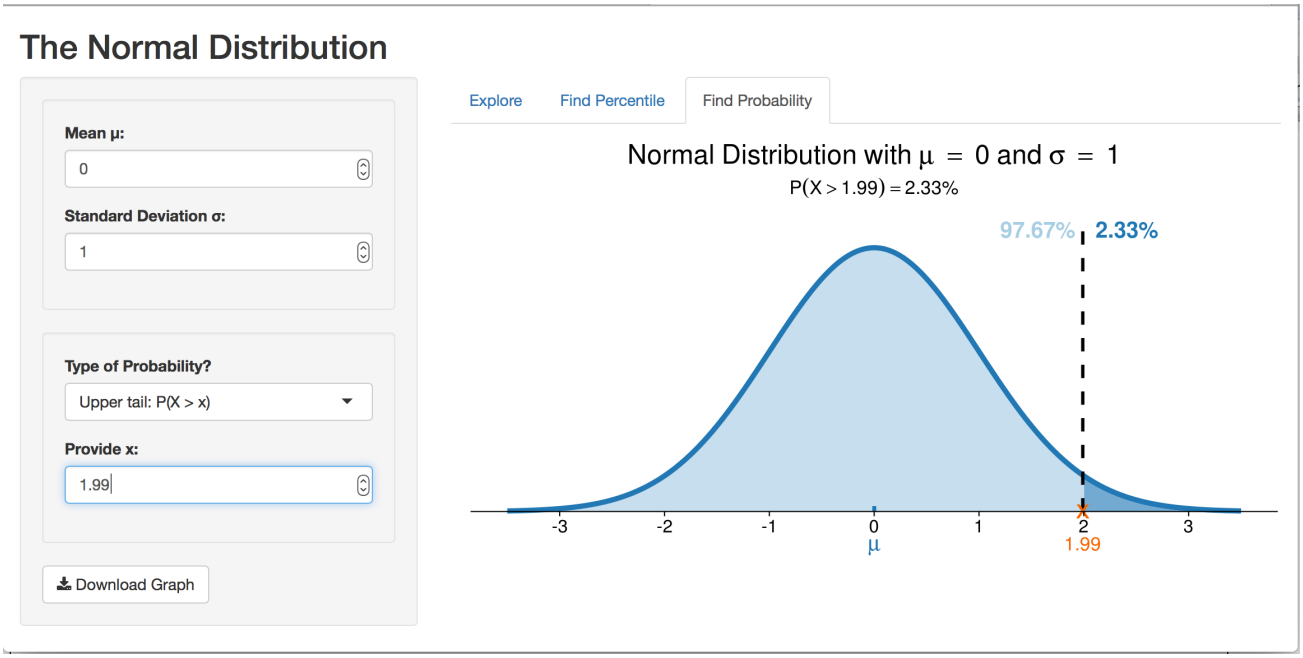
$$= \Pr[\text{Z-score} \geq 1.99] = .0233$$

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 darker 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 (the darker blue shaded area is a probability and is the p-value)



https://istats.shinyapps.io/NormalDist/ *Tip: **www.artofstat.com** > Online Web Apps > Normal Distribution > Click at top on tab FIND PROBABILITY*

Example of Statistical Hypothesis Test – Critical Region Approach

Source: BIOSTATS 540 Unit 8 Notes (Statistical Literacy) pp 19-23.

- We agree ***in advance (prior to collecting data)*** that we will honor a ***threshold test statistic value (this is what we mean by critical value)***, beyond which we will reject the null hypothesis. We'll do this even though, theoretically under the null hypothesis, such extreme values are still possible.

- This means that whenever we then do get a test statistic value that is beyond the critical value, ***if the null hypothesis is actually true***, then we will have ***incorrectly reject a true null hypothesis.*** Under these circumstances, we will have made a ***type I error***.

- ***How does this actually work?*** In developing a critical region test, we determine, ***before actual data collection***, the ***threshold statistic value*** (the "***critical value***") and the ***range of extreme values that lie beyond*** (this is called the ***critical region***) that will (in the future – after obtaining our data) prompt (rightly or wrongly!) rejection of the null hypothesis.

Example – again.
With standard care, cancer patients are expected to survive a mean duration of time equal to 38.3 months. Hypothesized is that a particular 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.

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

Null and alternative hypotheses

H_O: $\mu_{true} = \mu_O \leq 38.3$ **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:

- $X_1, X_2, ..., X_{100}$ 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.

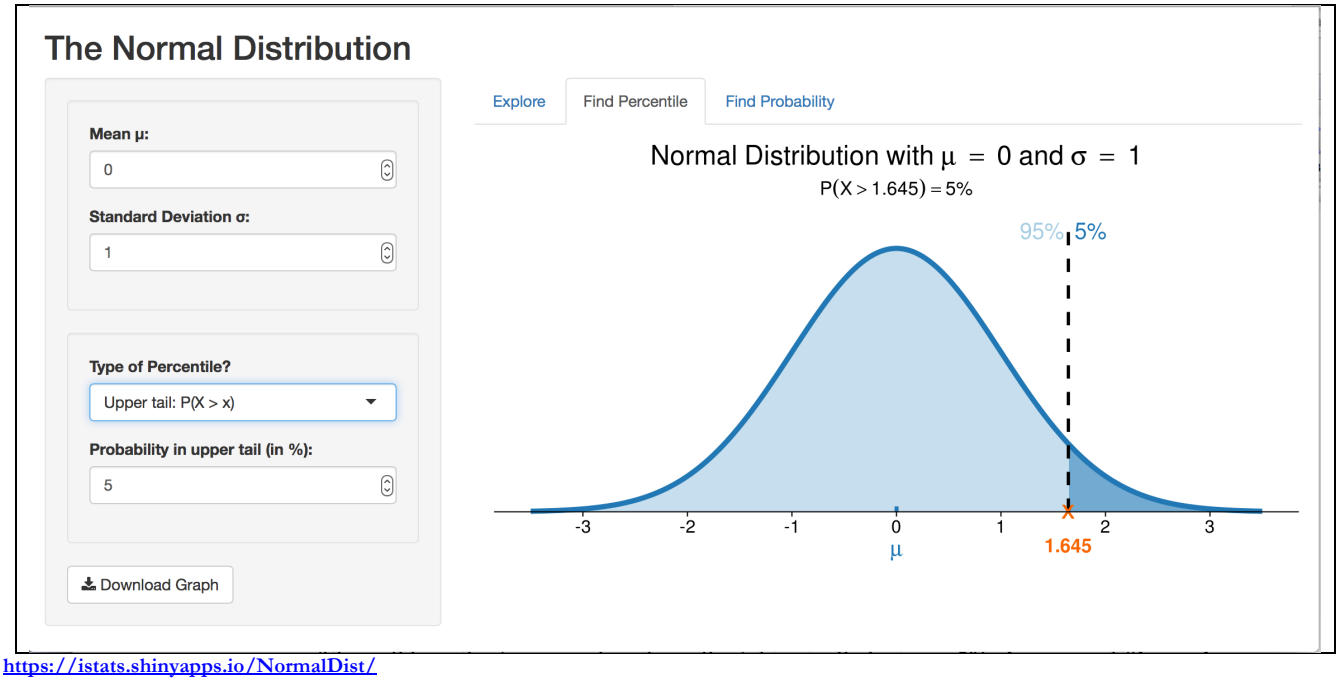
$$\text{Test Statistic} = \text{z-score} = \frac{\bar{X}_{n=100} - \mu_{null}}{SE(\bar{X}_{n=100})}$$

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

Step 1: **Consider the *direction* of the alternative, relative to the null hypothesis**
(eg – if the new treatment is better, survival will be longer):
In this example, the alternative hypothesis model is one sided. Moreover, we are looking to reject the null hypothesis in favor of the alternative (improved survival) if the test statistic is extremely **large.**

Step 2: **Solve for the threshold and critical region of the test statistic**
Here is where our 5% chances (see previous page) get used. Work with the null hypothesis model. Specifically, in this example, the **Z-statistic** (see bottom of page 33) will be distributed Normal(0,1) when the null hypothesis is true. We have already agreed that we want to solve for a threshold value of our Z-statistic, beyond which all values will prompt rejection of the null hypothesis, even when the null might actually be true. ***5% chances*** and ***"beyond which"*** translates to ***right tail area = .05***.

I used the link https://istats.shinyapps.io/NormalDist/
www.artofstat.com > Online Web Apps > Normal Distribution > Click at top on tab FIND PERCENTILE



https://istats.shinyapps.io/NormalDist/

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

How? We do this by setting the formula for the z-score equal to the value of the critical value for the z-score that was obtained in step 2, namely 1.6449.

$$\text{z-score} \geq 1.6449 \rightarrow$$

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

$$\bar{X}_{n=100} - \mu_{null} \geq (1.6449)*SE(\bar{X}_{n=100}) \rightarrow$$

$$\bar{X}_{n=100} \geq \left[(1.6449)*SE(\bar{X}_{n=100})\right] + \mu_{null} \rightarrow$$

$$\bar{X}_{n=100} \geq \left[(1.6449)*(4.33)\right] + 38.3 \rightarrow$$

$$\bar{X}_{n=100} \geq 45.42$$

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

Step 4: Interpret:

Going forward, after we have collected data and computed our test statistic, we will follow the following rule: If the observed $\bar{X}_{n=100}$ In has a value that is beyond the threshold value of 45.42, we will infer the alternative hypothesis, even though such extreme values are theoretically possible when the null hypothesis is true. That is – "this ***critical region*** one sided .05 test of the null versus alternative hypotheses has been defined to reject the null hypothesis for any $\bar{X}_{n=100} \geq 45.42$.

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

Now collect your data and compute your sample mean. In this example, the sample mean $\bar{X}_{n=100} = 46.9$ Because it is exceeds the threshold value of 45.42, it falls in the critical region.

Interpret.

Because the observed $\bar{X}_{n=100} = 46.9$ ***exceeds the value of the threshold 45.42*** and is, therefore, ***in the critical region***, in critical region parlance we say *"it is significant at the 0.05 level"*. → **reject the null hypothesis.** The conclusion is the same: these data provide statistically significant evidence, that compared to standard care, survival times on the new treatment are longer.

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: Lots, including …

- Treatment "A" is truly superior; ***or***

- Groups "A" and "B" were not comparable initially, rendering the "apparent" superiority of treatment "A" just a spurious one; ***or***

- Treatment "B" is actually superior, which means that the "apparent" superiority of treatment "A" is nothing more than having observed something very unlikely ("events of low probability do occur, just not very often").

… 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: BIOSTATS 540 Lecture Notes 9 (One Sample Inference), pp 9-10.

(2) One Sample, Normal Distribution
Test of μ (σ^2 NOT known)

Student t-test
For review, see: BIOSTATS 540 Lecture Notes 9 (One Sample Inference), pp 19-20.

(3) One Sample, Normal Distribution
Test of σ^2

Chi Square test
For review, see: BIOSTATS 540 Lecture Notes 9 (One Sample Inference), pp 25-27.

(4) One Sample Proportion, Binomial Distribution
Test of π

Z-test
For review, see: BIOSTATS 540 Lecture Notes 9 (One Sample Inference), pp 51-54.

(5) Paired Data Sample, Normal Distribution
Test of μ_d (σ^2_d known)

Z-test
For review, see: BIOSTATS 540 Lecture Notes 9 (One Sample Inference), pp 30-33.

(6) Paired Data Sample, Normal Distribution
Test of μ_d (σ^2_d NOT known)

Student t-test
For review, see: BIOSTATS 540 Lecture Notes 9 (One Sample Inference), pp 39-41.

(7) 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: BIOSTATS 540 Lecture Notes 10 (Two Sample Inference), pp 8-12.

(8) Two Independent Samples, Normal Distribution
Test of Equality of σ^2_1 and σ^2_2

F-test
For review, see: BIOSTATS 540 Lecture Notes 10 (Two Sample Inference), pp 24-26.

(9) Two Independent Proportions, Binomial Distributions
Test of $\pi_1 - \pi_2$

Z-test
For review, see: BIOSTATS 540 Lecture Notes 10 (Two Sample Inference), pp 32-34.