University of Hawai'i
Office of Biostatistics & Quantitative Health Sciences
John A. Burns School of Medicine

Biostatistics for Med Students

Lecture 2

John J. Chen, Ph.D.
Professor & Director of Biostatistics Core
UH JABSOM

JABSOM MD7

February 20, 2019

Lecture note: http://biostat.jabsom.hawaii.edu/Education/training.html

Lecture Objectives

- To understand basic research design principles and data presentation approaches
- To build a foundation which will facilitate the active participation in clinical research
- To fully grasp descriptive statistics
- To introduce key concepts of inferential statistics
- To survey some commonly used statistical approaches
- To be prepared for the USMLE Step 1 biostat/epi questions

Outline

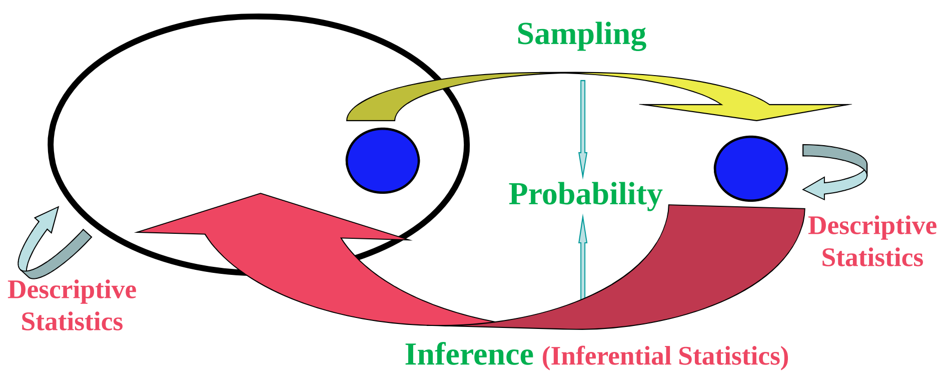
Lecture 1 (02/13/2019)

- The goal of statistics
- Introduction to descriptive biostatistics
- Basic research design principles and data presentation approaches

Lecture 2 (02/20/2019)

- Introduction to inferential statistics
- Commonly used statistical approaches

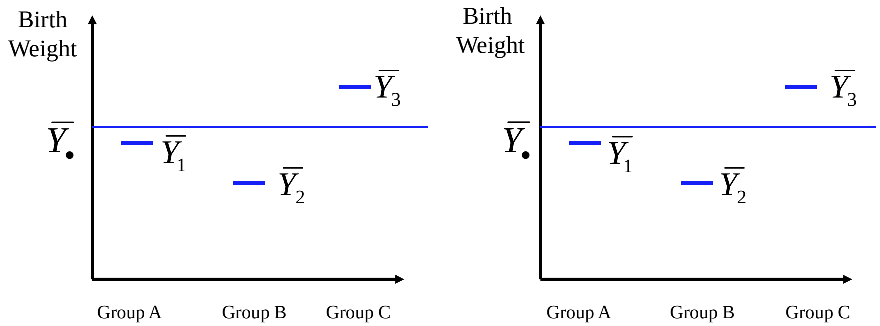
The Goal of Statistics



Population
[Parameters]
(μ, σ)

Sample
[Statistics]
($\bar{X}$, S)

Effects and Variability

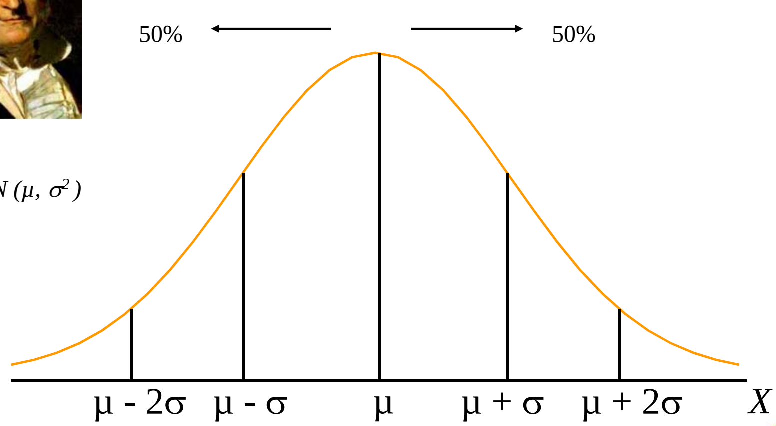


Note: Biological/clinical significance vs. statistical significance

The Normal Distribution

Carl Friedrich Gauss (1777-1855)

$X \sim N(\mu, \sigma^2)$



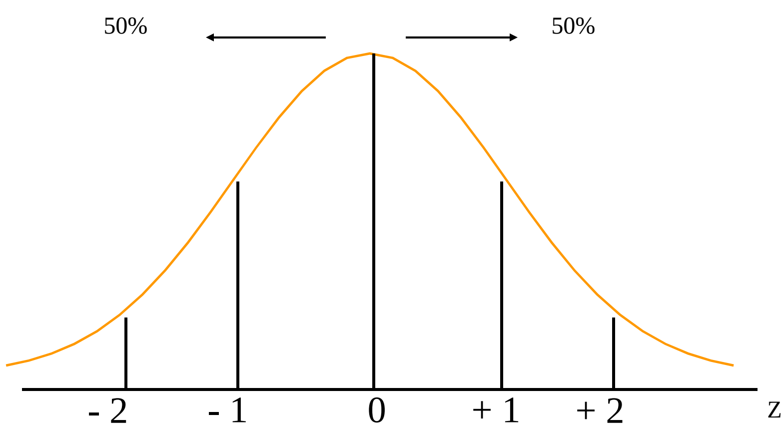
AUC For Normal Distribution

The Rule of Thumb:

Within one s.d.: 68.27% (2/3)
Within two s.d.: 95.45% (95%)
Within three s.d.: 99.74% (99%)

The Normal Distribution

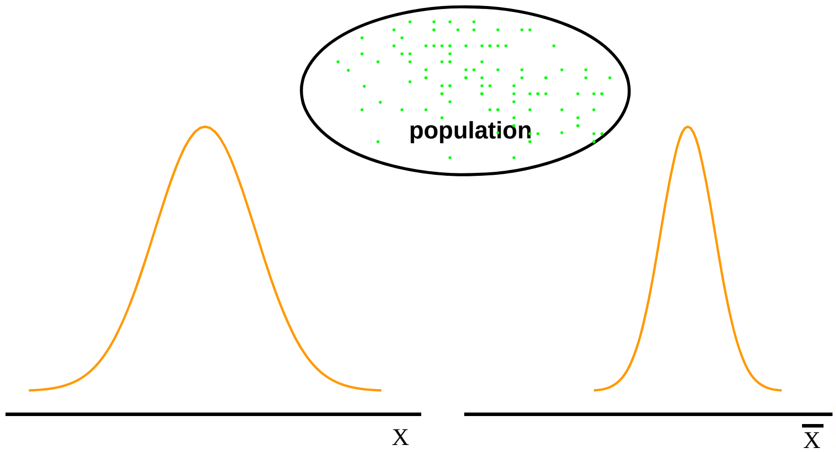
Standard normal distribution: $Z \sim N(\mu = 0, \sigma^2 = 1)$



Given $X \sim N(\mu, \sigma^2)$, we have $Z=(X - \mu)/ \sigma$.

Sampling Distribution

The distribution of individual observations versus the distribution of sample means:



Central Limit Theorem

The distribution of sample means (sampling distribution) from a population is approximately normal as long as the sample size is large, i.e.,

$$\bar{X} \sim N(\mu_{\bar{X}}, \sigma^2_{\bar{X}}) \quad \rightarrow \quad Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

1. The population distribution can be non-normal.
2. Given the population has mean μ, then the mean of the sampling distribution, $\mu_{\bar{X}} = \mu$.
3. If the population has variance σ^2, the standard deviation of the sampling distribution, or the standard error (a measure of the amount of sampling error) is $\sigma_{\bar{X}} = s.e.(\bar{X}) = \dfrac{\sigma}{\sqrt{n}}$.

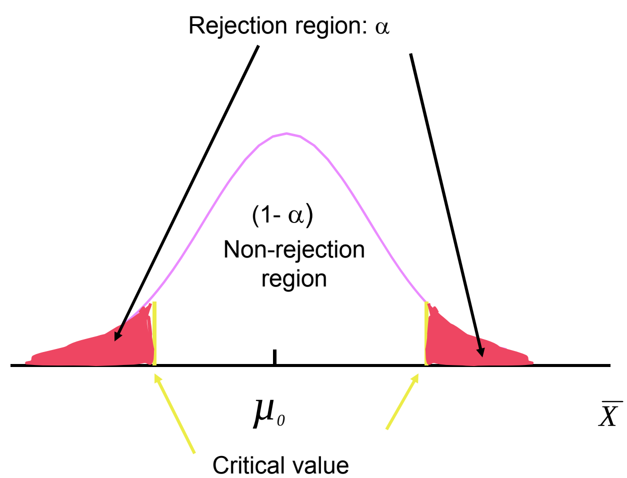
Hypothesis Testing

An Example:

Normal serum creatinine level depends on the population studied. From the literature a 4th year JABSOM med student found that one well-established study showed an average sCr of 0.56 (with a standard deviation of 0.15 mg/dL) for 2nd trimester Caucasian pregnant women living on the east coast. But based on her knowledge and experience, she believed that the μ of sCr among Japanese pregnant women in Hawaii seemed different.

She decided to test this by measuring sCr of 49 local Japanese 2nd trimester pregnant women.

Hypothesis Testing (cont.)



Hypothesis Testing

Basic steps of hypothesis testing:

1. State null (H_0:) and alternative (H_1:) hypotheses
2. Choose a significance level, α (usually 0.05 or 0.01)
3. Determine the critical (or rejection) region and the non-rejection region, based on the sampling distribution and under the null hypothesis
4. Based on the sample, calculate the test statistic and compare it with the critical values
5. Make a decision, and state the conclusion

Errors, Power, and Statistical Decision

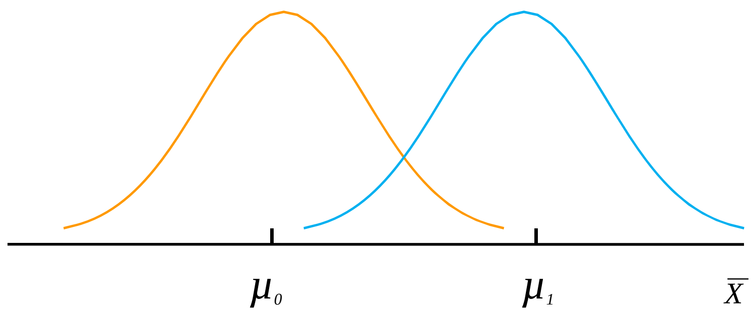
Type I Error (α)
- False positives, errors due to chance
- Reject H_0 when H_0 is true

Type II Error (β)
- False negatives
- Don't reject H_0 when H_1 is true

Power: $(1-\beta) = 1 - P$ (Type II Error)

		Truth: H_0 True	Truth: H_0 False
Decision	Reject H_0	α	$1-\beta$
	Not reject H_0	$1-\alpha$	β

Statistical Decision



p-values

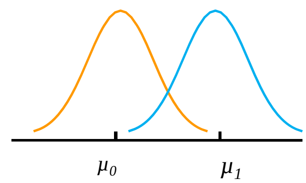
Interpretation:

The *p*-value is the probability of obtaining a result as extreme or more extreme than the one observed based on the current sample, given the null hypothesis is true.

Note: “Statistically significant” does not necessarily mean “biologically (or clinically) significant”!!!

Study Design: Power & Sample Size

Five General Design Factors:



DF 1. Effect difference
DF 2. Variability
DF 3. Statistical power (1- β)
DF 4. α level (Type I error)
DF 5. Sample size

$$\text{Sample Size} = f(DF1, DF2, DF3, DF4)$$

$$\text{Statistical Power} = f(DF1, DF2, DF4, DF5)$$

$$\text{Sample Size} = f(DF1 = 0.62 - 0.56, DF2 = 0.15, DF3 = 0.80, DF4 = 0.05) \approx 49.$$

Hypothesis Testing (cont.)

<u>Example (cont.):</u> Say, the average sCr of the sample of 49 locals is 0.60 mg/dL and the population standard deviation is 0.15 mg/dL (based on the literature).

Step 1. State H_0: and H_1:

$H_0 : \mu_{sCr} = 0.56$ vs. $H_1 : \mu_{sCr} \neq 0.56$

Step 2. Choose a significant level, say, α =0.05.

Step 3. Calculate the test statistic:

$$Z = \frac{\overline{X} - \mu_{sCr}}{\sigma / \sqrt{n}} = \frac{0.60 - 0.56}{0.15/\sqrt{49}} = 1.87.$$

Hypothesis Testing (cont.)

Step 4. Determine the critical region and the non-rejection region:

The critical value: $\pm$ *1.96.*

The rejection region: | Z | $\geq$ 1.96.

The non-rejection region: | Z | < 1.96.

Step 5. Make a decision, based on the sample, and state the conclusion: As the test statistic Z = 1.87 < 1.96, it is within the non-rejection region. Therefore, we do not reject the null hypothesis. We conclude that there is no evidence that the average sCr among local Japanese 2nd trimester women is different from 0.56 mg/dL.

Confidence Intervals

CIs for μ:

90% CI : $\bar{X} \pm 1.645 \frac{\sigma}{\sqrt{n}}$

95% CI : $\bar{X} \pm 1.960 \frac{\sigma}{\sqrt{n}}$

99% CI : $\bar{X} \pm 2.575 \frac{\sigma}{\sqrt{n}}$

Interpretation of 95% Confidence Interval for μ:

A. You can be 95% sure that the true mean (μ) will fall within the upper and lower bounds.

B. 95% of the intervals constructed using sample means, will contain the true mean (μ).

Guinness & The Student's *t*-Test

- A small sample from normal distribution
- Unknown population standard deviation, σ

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$ with *n -1* degrees of freedom.

The (Student's) t-distribution is very similar to normal distribution, with heavier tails.

t – Table (tail probabilities of the t-distributions)

Degrees of Freedom	2Q (Q)	0.10 (0.05)	0.05 (0.025)	0.01 (0.005)	0.005 (0.0025)	0.001 (0.0005)
1		6.3138	12.706	63.657	127.32	636.62
2		2.9200	4.3026	9.9251	14.0911	31.6075
3		2.3534	3.1825	5.8408	7.4533	12.9258
4		2.1318	2.7764	4.6040	5.5980	8.6087
5		2.0151	2.5706	4.0323	4.7734	6.8701
6		1.9432	2.4469	3.7075	4.3169	5.9590
7		1.8946	2.3646	3.4995	4.0293	5.4088
8		1.8595	2.3060	3.3555	3.8326	5.0421
9		1.8331	2.2621	3.2498	3.6895	4.7805
10		1.8125	2.2281	3.1693	3.5814	4.5871
11		1.7959	2.2010	3.1057	3.4967	4.4374
12		1.7823	2.1788	3.0545	3.4285	4.3184
13		1.7709	2.1604	3.0122	3.3726	4.2215
14		1.7613	2.1448	2.9768	3.3258	4.1412
15		1.7530	2.1314	2.9467	3.2862	4.0735
16		1.7459	2.1199	2.9207	3.2521	4.0157
17		1.7396	2.1098	2.8982	3.2226	3.9659
18		1.7341	2.1009	2.8784	3.1967	3.9224
19		1.7291	2.0930	2.8609	3.1738	3.8841
20		1.7247	2.0860	2.8453	3.1535	3.8502
21		1.7207	2.0796	2.8313	3.1353	3.8200
22		1.7171	2.0739	2.8187	3.1189	3.7928
23		1.7139	2.0687	2.8073	3.1041	3.7683
24		1.7109	2.0639	2.7969	3.0906	3.7461
25		1.7081	2.0595	2.7874	3.0783	3.7258
26		1.7056	2.0555	2.7787	3.0670	3.7073
27		1.7033	2.0518	2.7707	3.0566	3.6903
28		1.7011	2.0484	2.7632	3.0470	3.6746
29		1.6991	2.0452	2.7564	3.0382	3.6601
30		1.6973	2.0423	2.7500	[illegible]	[illegible]

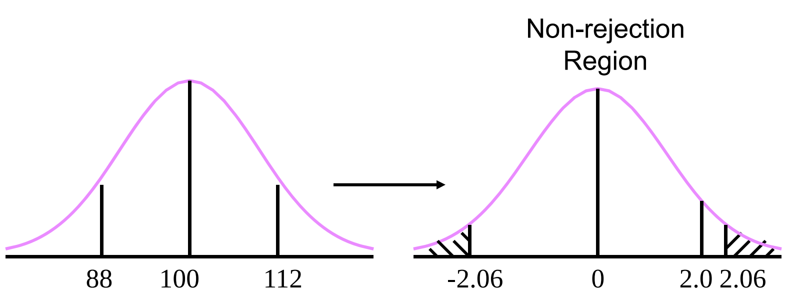
One Sample *t* – Test

<u>Problem:</u> Neonates gain (on average) 100 grams/wk in the first 4 weeks. A sample of 25 infants given a new nutrition formula gained 112 grams/wk (on average) with standard deviation = 30 grams. Is this statistically significant?

Test: H_0: $\mu = 100$ H_1: $\mu \neq 100$

One Sample *t* - Test

<u>Solution:</u>



$$t = \frac{112 - 100}{30 / \sqrt{25}} = 2.0 < t_{24,\,0.025} = 2.06$$

Therefore, do not reject (p-value=0.057).

Two Independent Sample *t* - Test

Population 1
(μ_1 , σ)

Sample 1
$(\bar{X}_1, s_1)$

Population 2
(μ_2 , σ)

Sample 2
$(\bar{X}_2, s_2)$

Test: H_0: $\mu_1 = \mu_2$ versus H_1: $\mu_1 \neq \mu_2$,
assuming $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Two Independent Sample *t* - Test

Problem: Two headache remedies

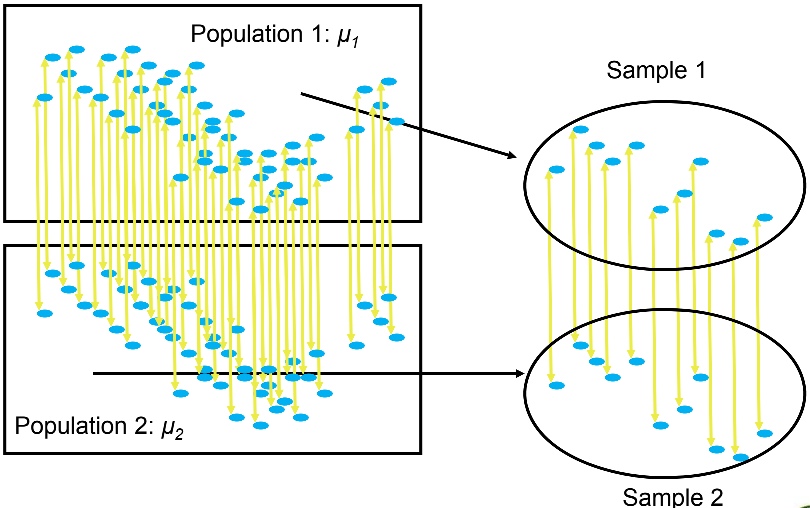
Brand A: $\bar{X}_1 = 20.1$, $s_1 = 8.7$, $n_1 = 12$

Brand B: $\bar{X}_2 = 18.9$, $s_2 = 7.5$, $n_2 = 13$

Test: H_0: $\mu_1 = \mu_2 \Rightarrow \mu_1 - \mu_2 = 0$

H_1: $\mu_1 \neq \mu_2$

Two Paired Sample *t* - Test



Test: H_0: $\mu_2 - \mu_1 = 0$ versus H_1: $\mu_2 - \mu_1 \neq 0$

Two Paired Sample *t* - Test

- Same subject for both treatments:
  -- placebo (X_1) versus active (X_2)
  -- before (X_1) versus after (X_2)
- Intra – individual comparison, e.g., left (X_1) versus right (X_2)

Approach: reduce data to one sample t-test problem. First, calculate the difference, $d = X_2 - X_1$, for each subject; then, perform one sample *t*-test on the *d* scores, with $d.f.=n-1$.

Two Paired Sample *t* - Test

Problem: Does the medication significantly lower blood pressure?

Subject	**Reaction to Placebo**	**Reaction to Med.**
1	**150**	**130**
2	**180**	**148**
3	**148**	**126**
4	**172**	**150**
5	**160**	**136**

Contingency Tables

The most common two-way tables: 2-by-2 tables

For example, smoking and lung cancer status

Lung cancer ← Response variable

Explanatory variable ↓ Smoking

Smoking \ Lung cancer	Yes	No	
Yes	n_{11}	n_{12}	$n_{1.}$
No	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	$n_{..}$

$$RR = \frac{n_{11}/n_{1\bullet}}{n_{21}/n_{2\bullet}}$$

$$RD = \frac{n_{11}}{n_{1\bullet}} - \frac{n_{21}}{n_{2\bullet}}$$

$$OR = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

Measures of association
- RD: risk difference (prospective studies)
- RR: relative risk (prospective studies)
- OR: odds ratio (prospective or retrospective)

Interpretations

Interpretation of estimates of RR/OR:

- If RR/OR is above 1, we say there is a positive relationship/association between risk factor and outcome.
- If RR/OR is below 1, we say there is a negative relationship/association between risk factor and outcome.

Interpretation of CI of population RR/OR:

- If the entire interval is above 1, we conclude that the probability that having the outcome of interest is higher in those with risk factor.
- If the entire interval is below 1, we conclude that the probability is lower in those with risk factor.
- We conclude that there is no evidence that the probabilities of the outcome of interest are different between those with vs. without risk factor, if the interval contains 1.

χ^2 Tests

Expected and observed frequencies are compared

- Goodness of fit of a single variable
- Test of independence of two variables

χ^2 Test of Independence

Observed:

	C1	C2	
R1	A	C	A+C
R2	B	D	B+D
	A+B	C+D	A+B+C+D

Expected:

$$\text{Exp} = \frac{\text{(row total) * (column total)}}{\text{(grand total)}}$$

e.g., $E_{1,1}$=(A+C)*(A+B) / (A+B+C+D)

$$\chi^2 = \sum_{i,j} \frac{(Obs_{ij} - Exp_{ij})^2}{Exp_{ij}}$$

d.f. = (r-1)(c-1)=(2-1)*(2-1) = 1*

Fisher's Exact Tests

<u>Fisher's Exact Test</u> – for small expected frequencies
If any expected frequencies are < 2 or if half of the expected frequencies are < 5, you should use Fisher's Exact Test instead of χ^2.

Fisher's Exact Tests

Fisher's Tea Tasting Experiment

		Poured First (Guessed)		
		Milk	*Tea*	
Poured First (Actual)	*Milk*			*4*
	Tea			*4*
		4	*4*	*8*

Fisher's Exact Tests

Fisher's Tea Tasting Experiment (cont.):

Based on hypergeometric distribution, the p-value is the sum of all probabilities for tables that give even more evidence in favor of the lady's claim.

		Poured First (Guess)		
		Milk	*Tea*	
Poured First (Actual)	*Milk*	0	4	
	Tea	4	0	

… …

	Poured First (Guess)		
	Milk	*Tea*	
Milk	3	1	
Tea	1	3	

	Poured First (Guess)		
	Milk	*Tea*	
Milk	4	0	
Tea	0	4	

p-value = $P_{(1,1)}(3) + P_{(1,1)}(4) = 0.229 + 0.014 = 0.243$

Therefore, the experiment did not establish a significant association between the actual order of pouring and the woman's guess.

Review: Lecture Objectives

- To understand basic research design principles and data presentation approaches
- To build a foundation which will facilitate the active participation in clinical research
- To fully grasp descriptive statistics
- To introduce key concepts of inferential statistics
- To survey some commonly used statistical approaches
- To be prepared for the USMLE Step 1 biostat/epi questions