Statistical analysis for NAPPA

Yunro Chung, Ph.D.
Assistant Professor
Arizona State University

Overview

Large population → Samples → Experiment(s) → Data Analysis

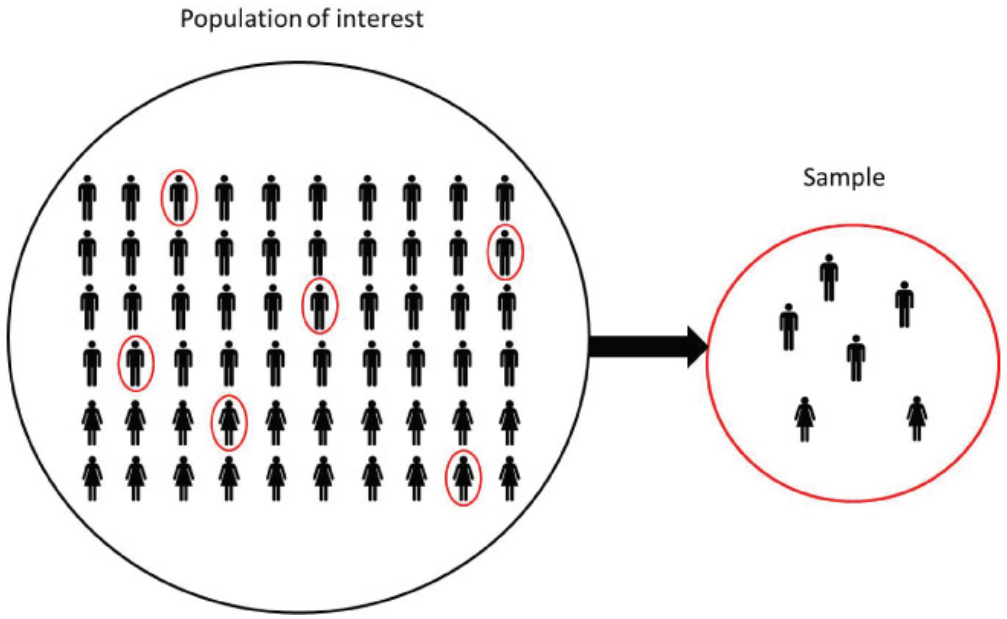
In practice, we observe a single sample.

What if we had drawn a different sample and gotten different result?

To answer this question, we need the form of a formal theoretical framework.

Statistical Inference:
How to draw scientific conclusions about a population using samples drawn?

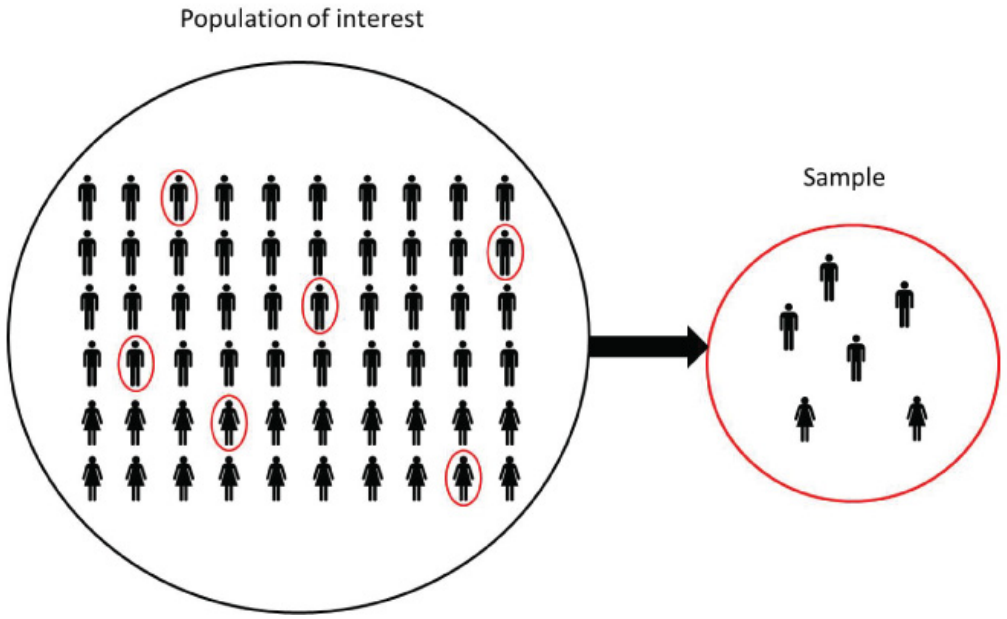
Population vs Sample



Population [N] **Sample [n]**

- **(Target) Population** [N] is all members of a specified group which an investigator wishes to draw a conclusion
  - Example: Breast cancer patients in the USA
- **Sample** [n] is a part of a population used to describe the whole group, i.e. subset of population
  - Example: Breast cancer patients in AZ

Parameter and Statistics



Population [N]

μ: population mean

σ: population standard deviation

Sample [n]

$\bar{X}$: sample mean

s: sample variance

- **Parameter** is a descriptive measure computed from the data of a population, denoted by lower case Greek letters, e.g., μ, σ
- **Statistics** is a descriptive measure computed from the data of a sample, denoted by Latin letters, e.g., $\bar{X}, s$

Exploratory data analysis

Mean & Standard Deviation

- Mean = sum of values divided by the number of values

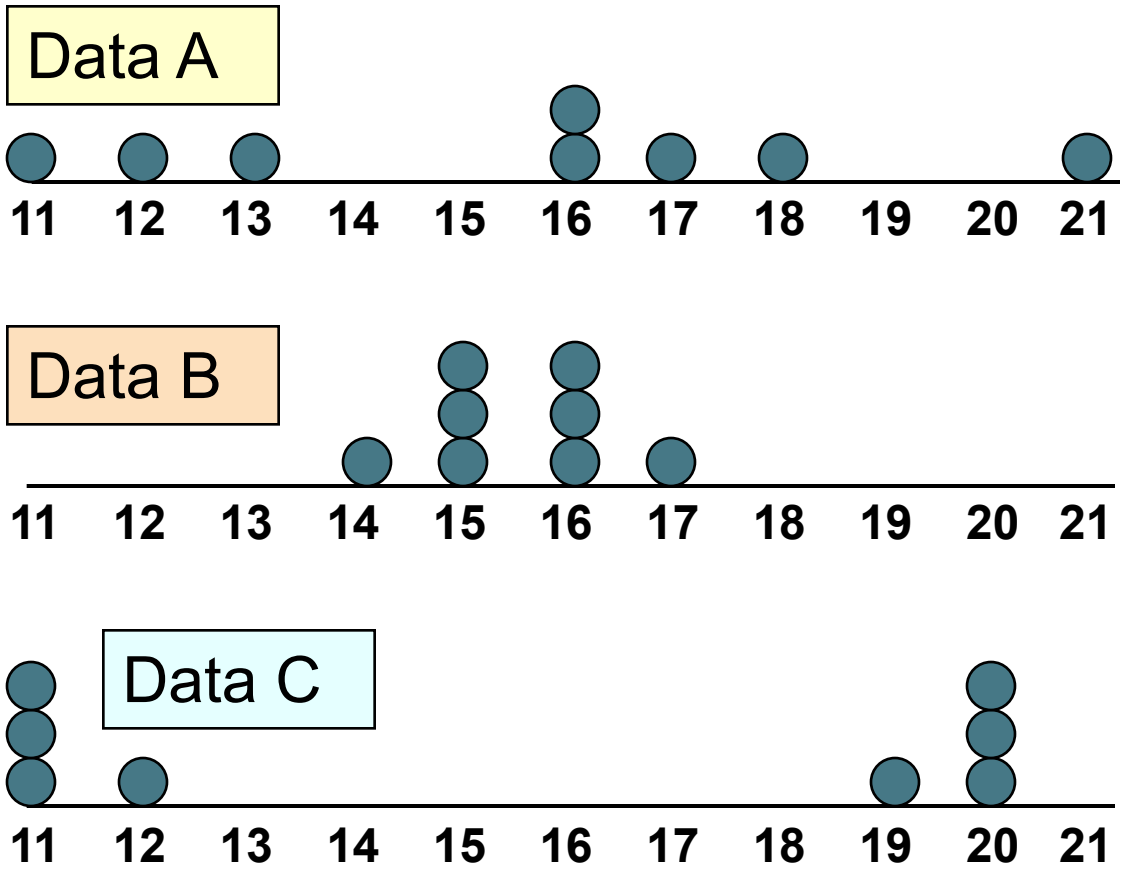
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- Standard Deviation.
  - Shows variation about the mean.
  - The larger the standard deviation, the larger the variability.

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

- Variance=s^2.

Comparing Standard Deviations



Mean = 15.5
s = 3.338

Mean = 15.5
s = .9258

Mean = 15.5
s = 4.57

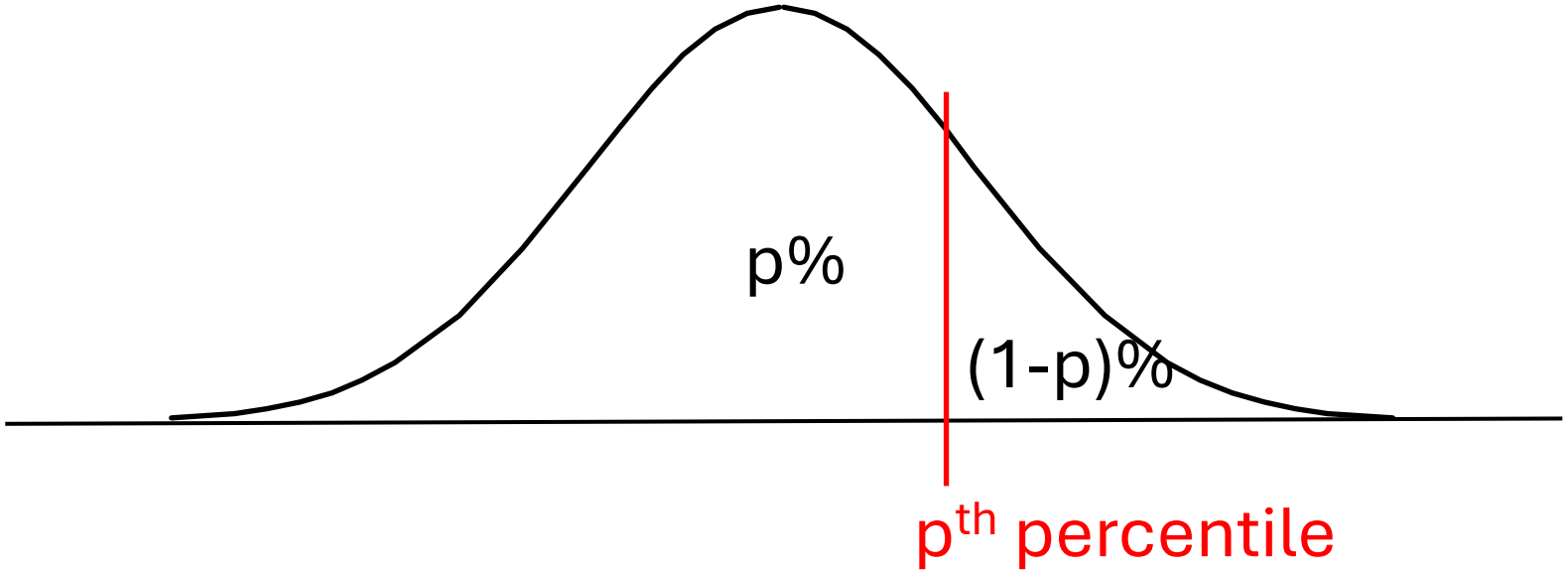
Median & IQR

- Mean and standard deviation are affected by extreme values (or outliers).

- Median is the **middle point** of a set of n values.
  - If n is odd, the median is the middle number
  - If n is even, the median is the average of the two middle numbers

Percentiles

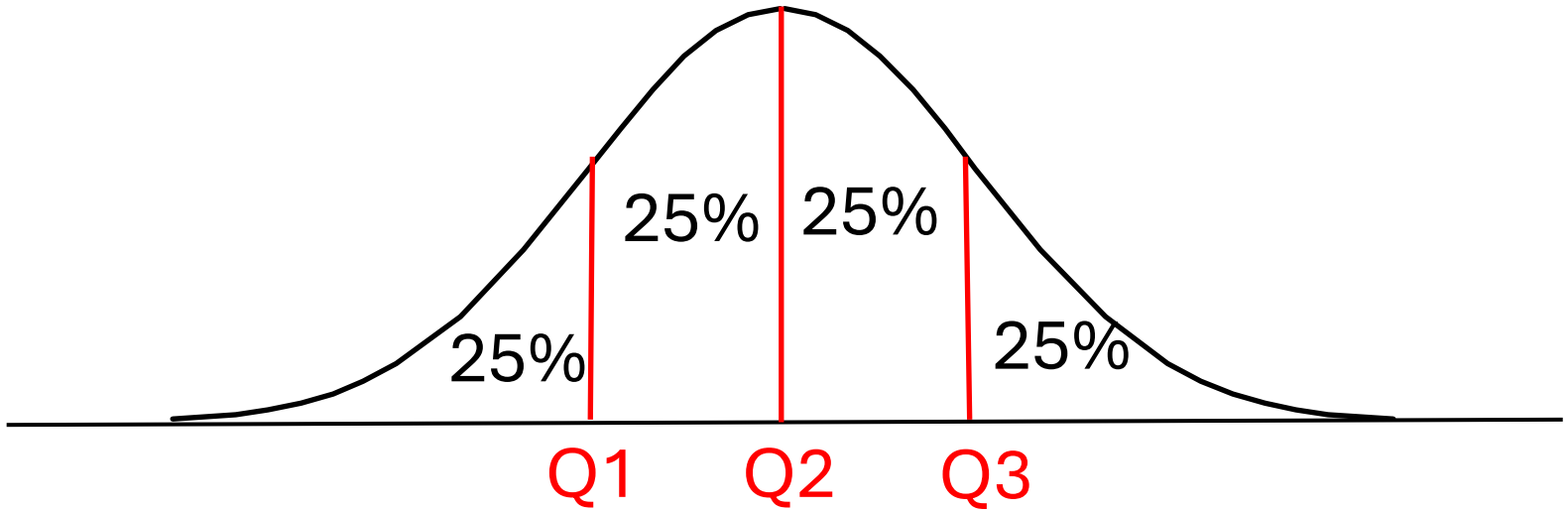
- The p^{th} percentile is the value below which p% of the observations fall.



- Example: 70^{th} percentile indicates
  - 70% are less than or equal to the 70th percentile
  - (100 – 70)%=30% are greater than or equal to 70th percentile

Quartiles

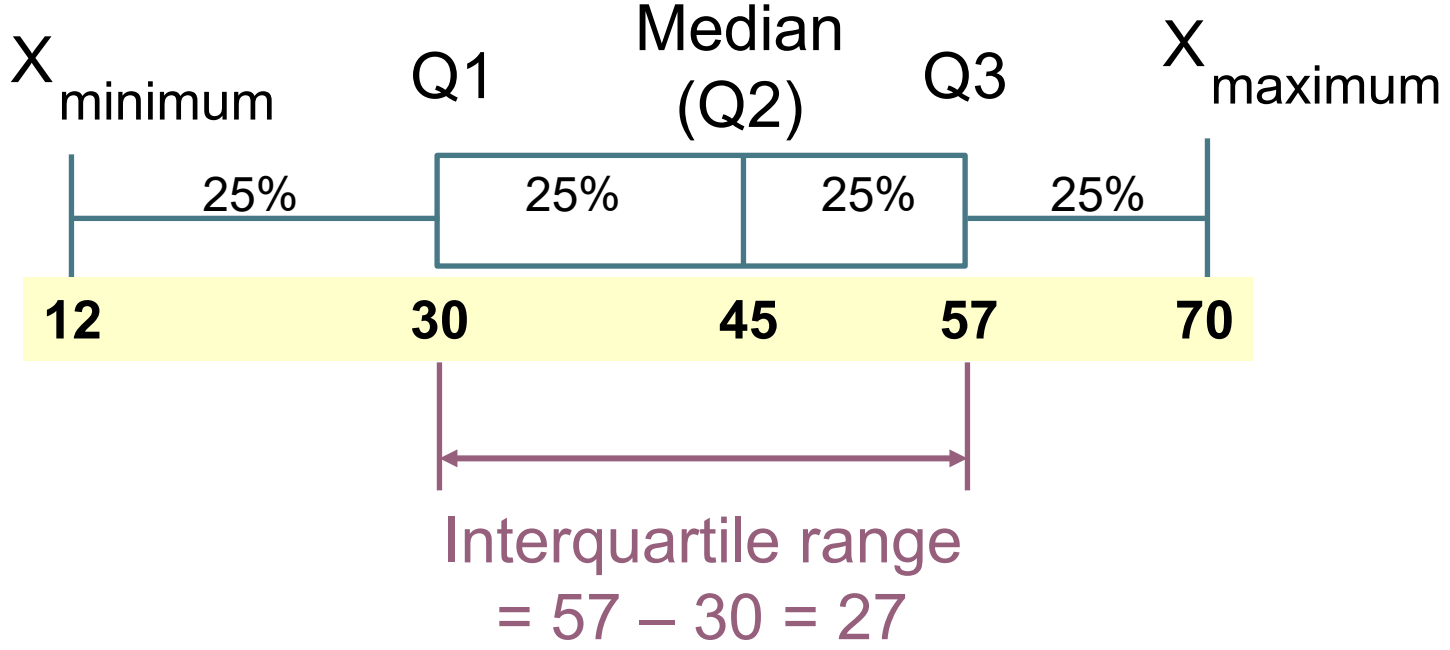
- 1st quartile (Q1) = 25th percentile
- 2nd quartile (Q2)= 50th percentile= median
- 3rd quartile (Q3) = 75th percentile



- Quartiles split the ranked data into 4 equal groups

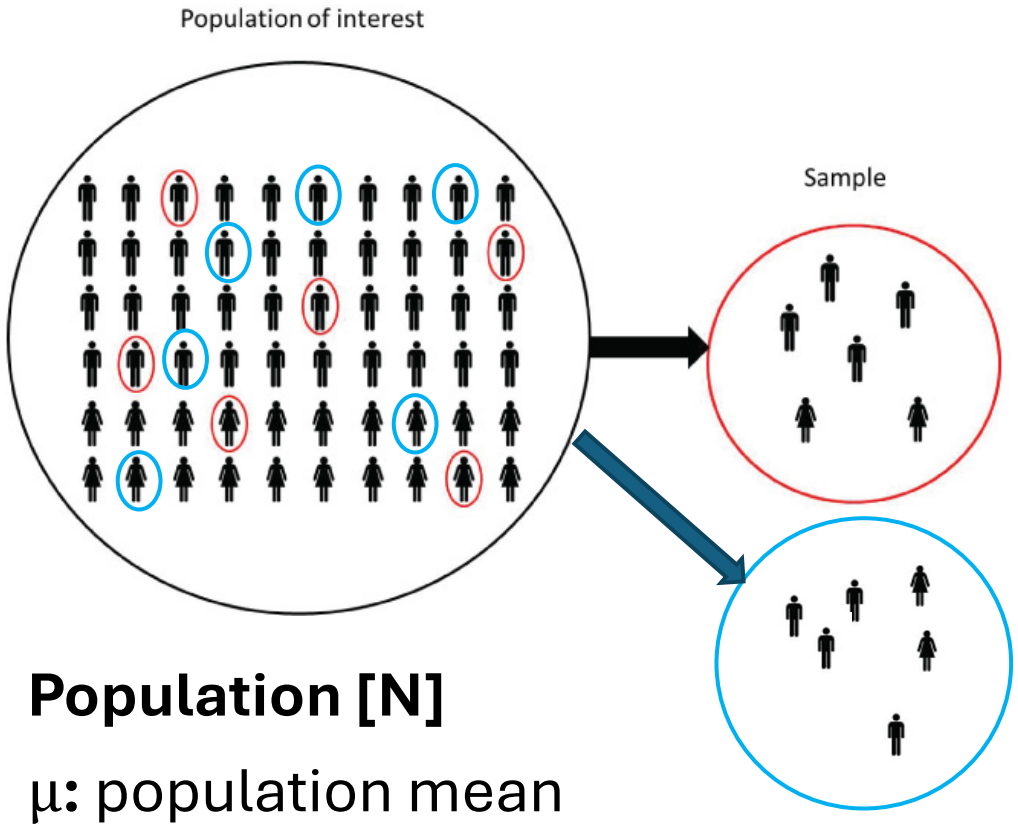
Inter-Quartile Range (IQR)

- IQR: The difference between the 75th percentile (often called Q3) and the 25th percentile (Q1). The formula for interquartile range is therefore: IQR=Q3-Q1.
- Range of the middle 50% of values



Sampling distribution of the mean

Sampling



Population [N]

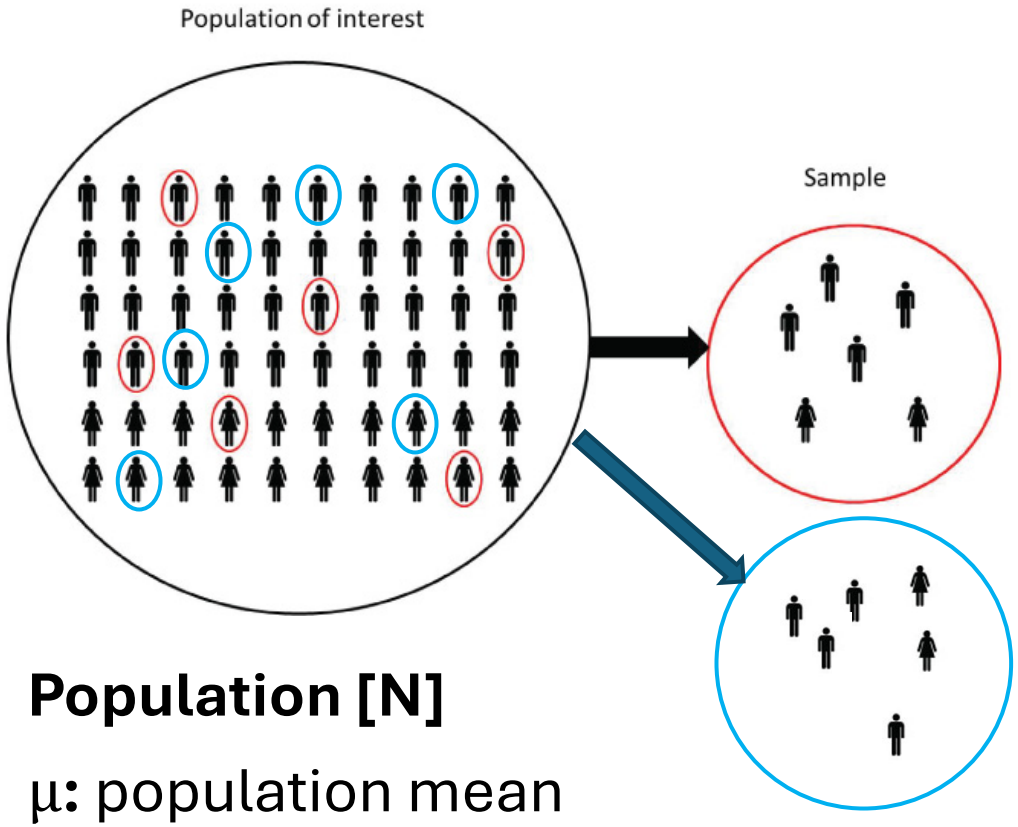
μ: population mean

In many cases, we are interested in μ

Problem:

- We are not able to examine all members in the population.
- $\overline{X}$s from samples 1 and 2 could be different even though the samples are drawn randomly.

Sampling distribution of means



Population [N]

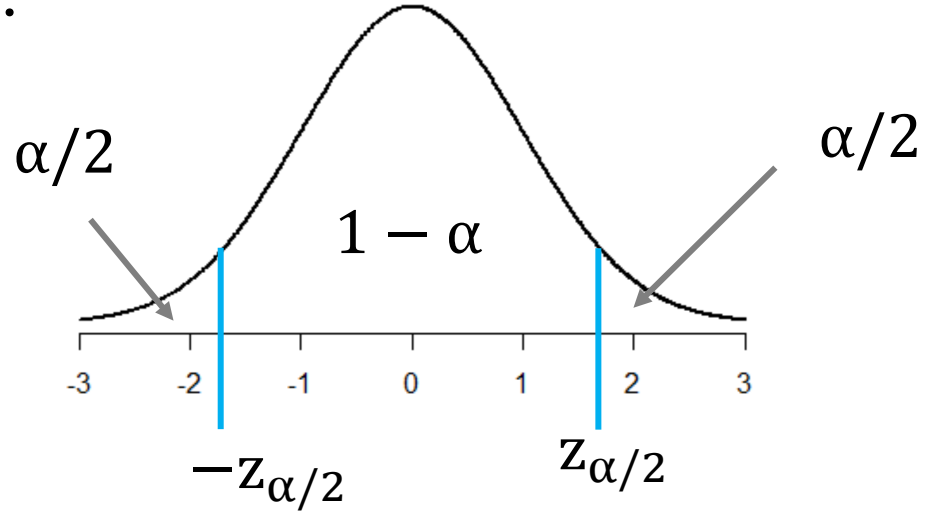
μ: population mean

Approach:

- Suppose that we hypothetically repeat sampling, i.e., samples 1, 2, 3, … and denote corresponding sample means as $\overline{X}_1$, $\overline{X}_2$, $\overline{X}_3$, …
- $\overline{X}$ can be viewed as a *random variable* with possible outcomes $\overline{X}_1$, $\overline{X}_2$, $\overline{X}_3$, …
- Can we assign a probability to the $\overline{X}$ you got from your sample?
- Yes. The probability distribution of $\overline{X}$ is called sampling distribution of the mean of size n.

The central limit theorem

- Given a population with mean μ and standard deviation σ and n is large enough, the distribution of the sample means will be approximately normally distributed.
- $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is normally distributed with mean 0 and standard deviation 1 if n is large enough.
- We write $Z \rightarrow N(0,1)$ as $n \uparrow \infty$.
- $\sigma/\sqrt{n}$ is the standard error of the mean (SEM).



Confidence Intervals

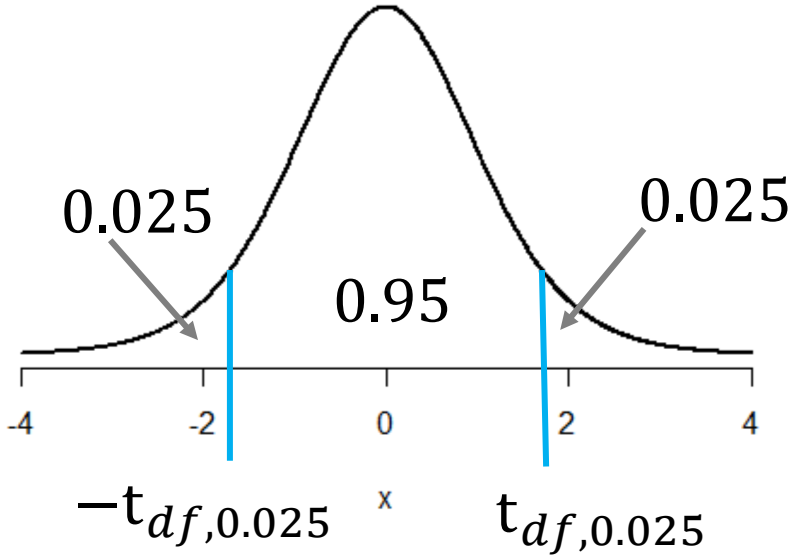
Point vs Interval Estimations

- Point estimation
  - μ is estimated by $\overline{X}$ from the sample.
  - It does not provide any information about the variability.
  - We do not know how close $\overline{X}$ to μ?

- Interval estimation
  - It provides a range of reasonable values that are intended to contain μ.
  - The range of values is called a confidence interval.

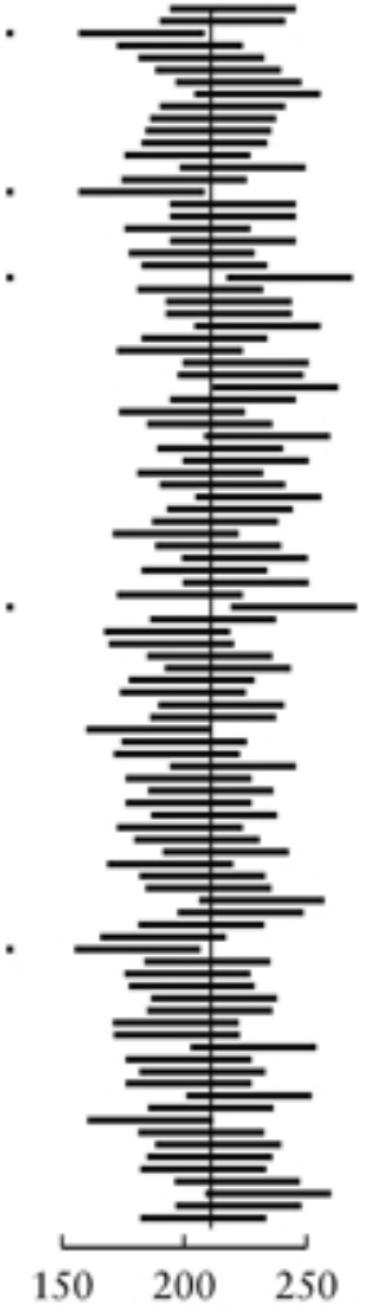
Confidence Intervals

- (1-α)x100% CI of μ is ($\overline{X}$-$z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$,$\overline{X}$+$z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$), if σ is unknown.
- Confidence level (1 − α): The probability that the confidence interval contains the true population parameter.
- In practice, σ is unknown and replaced with s.
- 95% confidence interval of μ is ($\overline{X}$-$t_{df,0.025}\frac{s}{\sqrt{n}}$,$\overline{X}$+$t_{df,0.025}\frac{s}{\sqrt{n}}$), where df (degree of freedom)=n-1



Confidence Intervals

- Example. Consider the distribution of serum cholesterol levels for all males in the United States who are hypertensive and who smoke. We draw a sample of size 100 from the population of hypertensive smokers and that these men have a mean serum cholesterol level of $\bar{X}$=217mg/100ml (95% CI: 191, 243).
- Suppose that that μ = 211, which is unknown.
- **How to interpret?**
  - We do not say that there is a 95% probability that μ lies between these values, since μ is fixed, not random.
  - If we were to draw 100 random samples of size 100 from this population and use each one to construct a 95% confidence interval we would expect that, on average, 95 of the intervals would cover the true population mean μ = 211 and 5 would not.

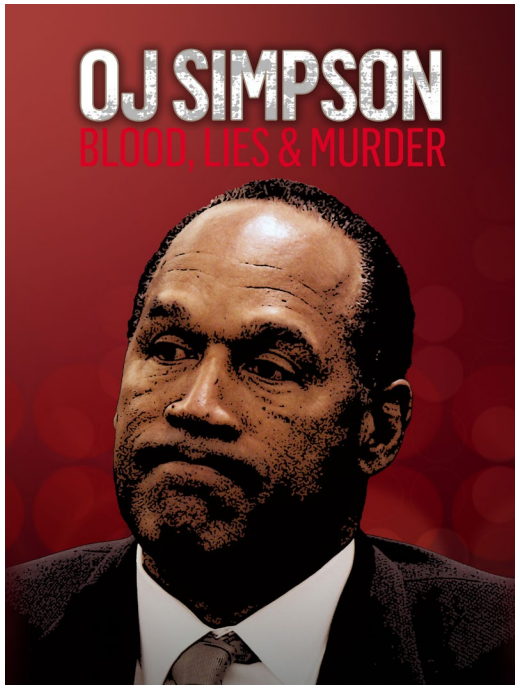


Hypothesis testing

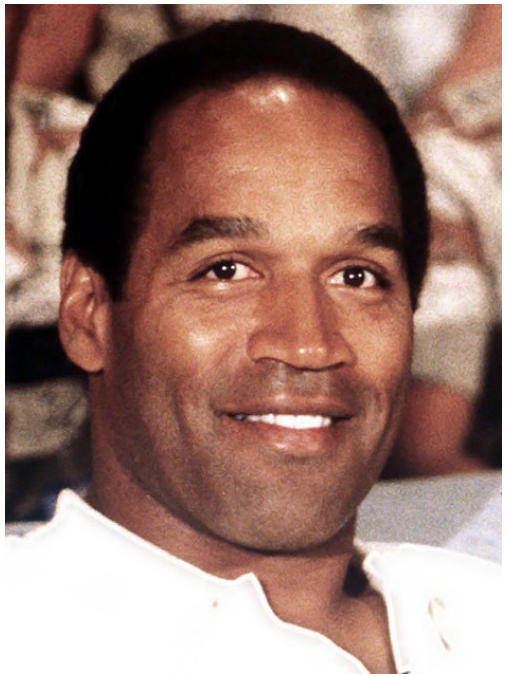
The 1995 OJ Simpson trial

- Nicole Brown Simpson and Ronald Gold man were founded brutally murdered on June 12, 1994.
- Nicole was the wife of O.J. Simpson, former NFL player (winner of 1968 Heisman Trophy, Hall of Famer).
- Evidence collected:
  - bloody size 12 Bruno Magli shoeprint, bloody glove, blood spots on white Ford Bronco, etc.

○ *O.J. Simpson was accused of the murders but assumed to be innocent by law.*



The O.J. Simpson Story (1995)



The 1995 OJ Simpson trial

- The individual on trial was *either innocent (H_o) or guilty (H_A)* but *assumed to be innocent by the law*, i.e. assume H_o is true.

Verdict of Jury	Innocent (H_o)	Guilt (H_A)
Guilty	Incorrect	OK
Not Guilty	OK	Incorrect

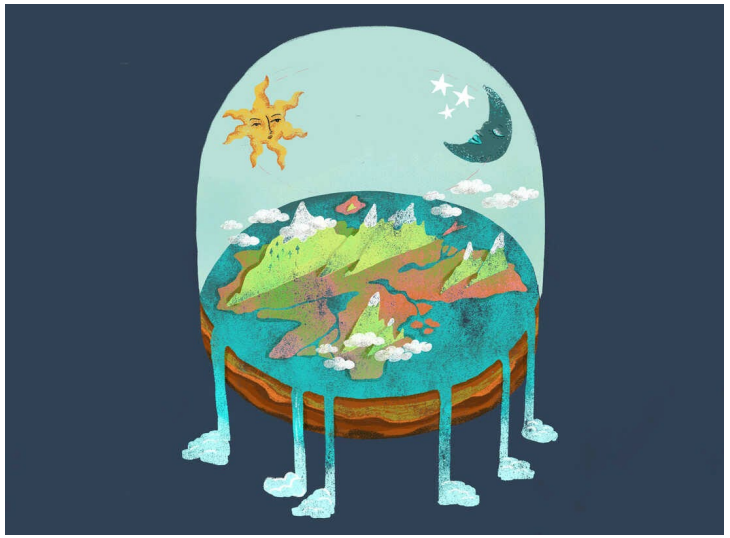
- After evidence pertaining to the case has been presented, the jury finds the defendant
  - guilty if there is sufficient amount of evidence (*reject the H_o*).
  - not guilt if there is no sufficient amount of evidence (*fail to reject the H_o*).
- OJ Simpson was not found guilt. ***Did it prove he was innocent?***
  - *There may be no sufficient evidence.*

Judicial Analogy to Hypothesis Testing

- H_0: no difference, no effect, no relationship vs H_A: not H_0

  e.g. Earth is flat vs Earth is round

- *Assume the H_o is true.*



Decision	$\mathbf{H_o}$ **is true**	$\mathbf{H_A}$ **is true**
Reject $\mathbf{H_o}$	Error (type I error)	Correct
Do not reject $\mathbf{H_o}$	Correct	Error (type II error)

- After collecting and analyzing data, its result supports
  - *rejection of the H_o .*
  - *failure to reject the H_o .*
- Fail to reject the H_o does not prove H_o is true.
  - *We never say we accept the H_o .*

- Type I error (reject H_o when H_o is true)
- Type II error (fail to reject H_o when H_A is true)

Terminology

Result of Test	H_0 is true (Innocent)	H_A is true (Guilty)
Reject H_0 (Guilty)	Incorrect (Type I error, α)	Correct (Power, 1-β)
Do not reject H_0 (Not guilty)	Correct	Incorrect (Type II error, β)

- *As you decrease α, you increase β and vice versa.*
- *General strategy is to fix α=5% and try to minimize β (or maximize 1-β).*

- α (commonly referred to as level of significant): probability of making a type I error.
- β: probability of making a type II error.
- 1-β: (commonly referred to as power): probability of reject H_0 when H_A is true

- In most applications, α=0.05.
- Power increases as n increases.

One Sample Test

One sample Z-test or T-test

- Test to see if the mean of a population (μ) differs from a known value (μ_0)
  - H_o: μ=μ_0 vs H_A: μ≠μ_0
- We do not test H_o: $\overline{X}$=μ_0 vs H_A: $\overline{X}$≠μ_0, i.e., we make a conclusion on the population parameter μ, not the sample mean $\overline{X}$.
- σ is known → Z-test
- σ is unknown → T-test

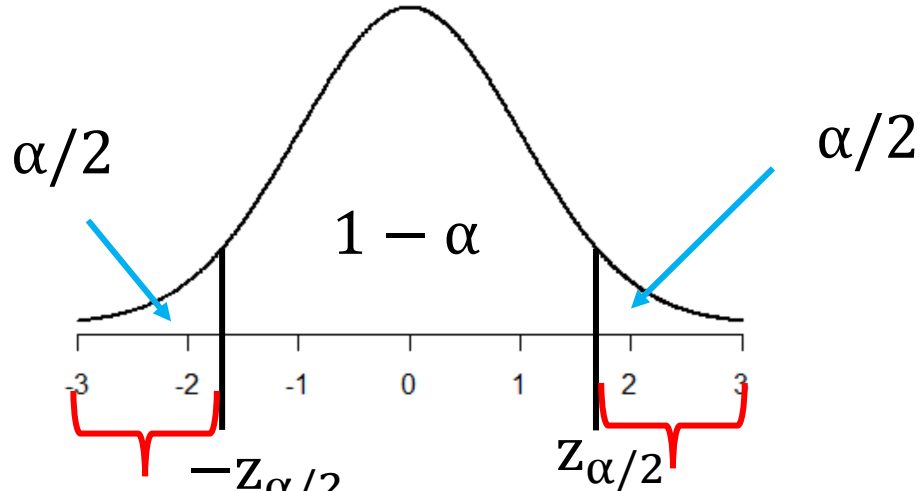
One sample Z-test

- H_o: $\mu=\mu_0$ vs H_A: $\mu\neq\mu_0$
- Under the null hypothesis,
  - $Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$ has an approximate normally distributed with mean 0 and standard deviation 1, N(0,1), if n is sufficiently large.
  - Z is called a test statistics, which can be calculated from a sample.
  - μ_0 is a constant that researchers can set, e.g., μ_0=211 mg/100 ml.

  - A smaller |Z| value would more support the H_o.
  - A larger |Z| value would more support the H_A.

One sample Z-test

- Under the H_o, Z is approximately N(0,1) with a large n.



Rejection region *Rejection region*

α	$z_{\alpha/2}$
0.01	2.576
0.05	1.960
0.10	1.645

- $z_{\alpha/2}$ and $-z_{\alpha/2}$ are called critical values.
- If the test statistics is in the rejection region, we reject the H_o.
- Otherwise, we fail to reject the H_o.

One sample Z-test

- We reject H_o if
  1. test statistics is in rejection region.
  2. Equivalently, p-value is less than or equal to α, e.g., 5%.
     - ✓The p-value is the probability that the test statistic is in the rejection region given H_o is true.
  - It implies that we reject incorrectly 5% of the time.
  - That is, given many repeated tests of significance, 5 times out of 100 we will erroneously reject H_o when it is true.

One Sample t-test.

- H_o: $\mu=\mu_0$ vs H_A: $\mu\neq\mu_0$
- Z test is not commonly used in practice because σ is typically unknown, i.e. $Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$ cannot be computed without knowing σ.
- By replacing σ with s, $t = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$ has a t distribution with n-1 degrees of freedom (df) under the null hypothesis.
- We then reject the null if the t value is in rejection region or corresponding p-value is less than α.

Two-sample t-test

Two-sample t-test

- Compare means between two independent groups.

		Group 1	Group 2
Population	Mean	μ_1	μ_2
	Standard deviation	σ_1	σ_2
Sample	Mean	$\overline{x}_1$	$\overline{x}_2$
	Standard deviation	s_1	s_2
	Sample size	n_1	n_2

1. State the hypotheses: H_o: $\mu_1=\mu_2$ vs H_A: $\mu_1\neq\mu_2$
2. Compute test statistic: $t = \frac{(\bar{X}_1-\bar{X}_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$
3. Compare p-value with α.

Two-Sample t-Test

- Two groups variances can be equal or not equal.
- Under the null hypothesis,
  - Equal variance:
  - $t = \frac{(\bar{X}_1 - \bar{X}_2)}{s_p\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ has a t-distribution with degree of freedom $n_1 + n_2 - 2$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$ (pooled variance)

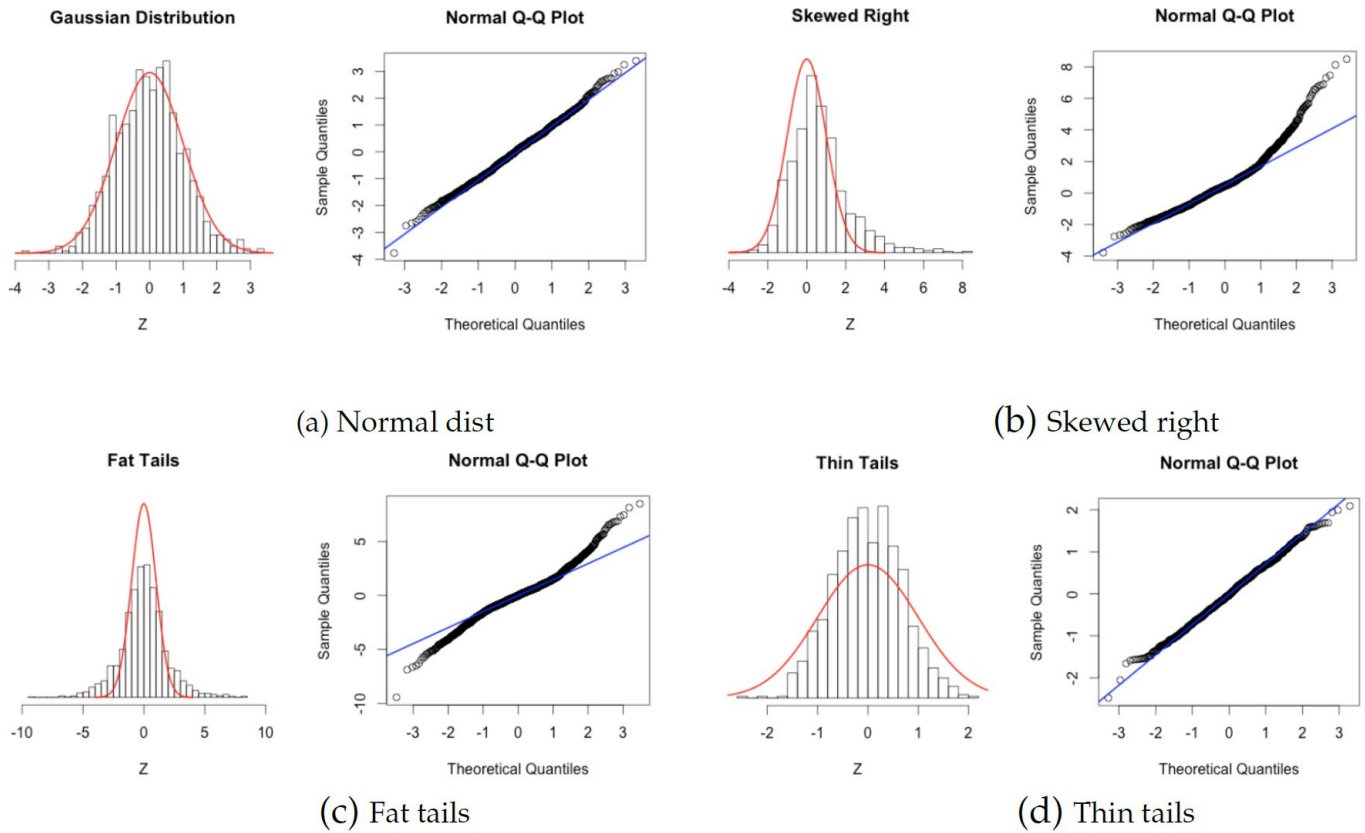
  - Unequal variance:
  - $t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$ has a t-distribution with the degree of freedom

$$\frac{[(S_1^2/n_1) + (S_2^2/n_2)]^2}{[(S_1^2/n_1)^2/(n_1 - 1) + (S_2^2/n_2)^2/(n_2 - 1)]}.$$

Normality Assumptions

- The Z or T tests can be applied when a) n is large or b) data is normally distributed.
- How to check the normality?
- Histogram, normal quantile-quantile (Q-Q) plot, Shapiro-Wilk test, etc.

* Q-Q Plots:



(a) Normal dist (b) Skewed right

(c) Fat tails (d) Thin tails

Nonparametric test

Wilcoxon Rank-Sum Test

- AKA: Mann-Whitney U test; Wilcoxon-Mann-Whitney test
- Wilcoxon rank-sum test is a nonparametric alternative as two-sample t-test.
- Key ideas:
    1. Temporarily combine the two samples into one big sample, then replace each sample value with its rank
    2. Find the sum of the ranks for either one of the two samples
    3. Compare the rank sums between the two groups.

Wilcoxon Rank-Sum Test

- BMI measurements between men and women.
- Numbers in parentheses are their ranks beginning with a rank of 1 assigned to the lowest value of 17.7.
- Tied observations are assigned an average rank
- R_1 and R_2: sum of ranks
  - If the null is true, R_1 and R_2 would be similar.
  - We reject the null if R_1 and R_2 are different.
- H_o: The population medians are the same between two groups vs H_1: Not H_o
- P-value is computed based on R_1, R_2 , n_1, n_2

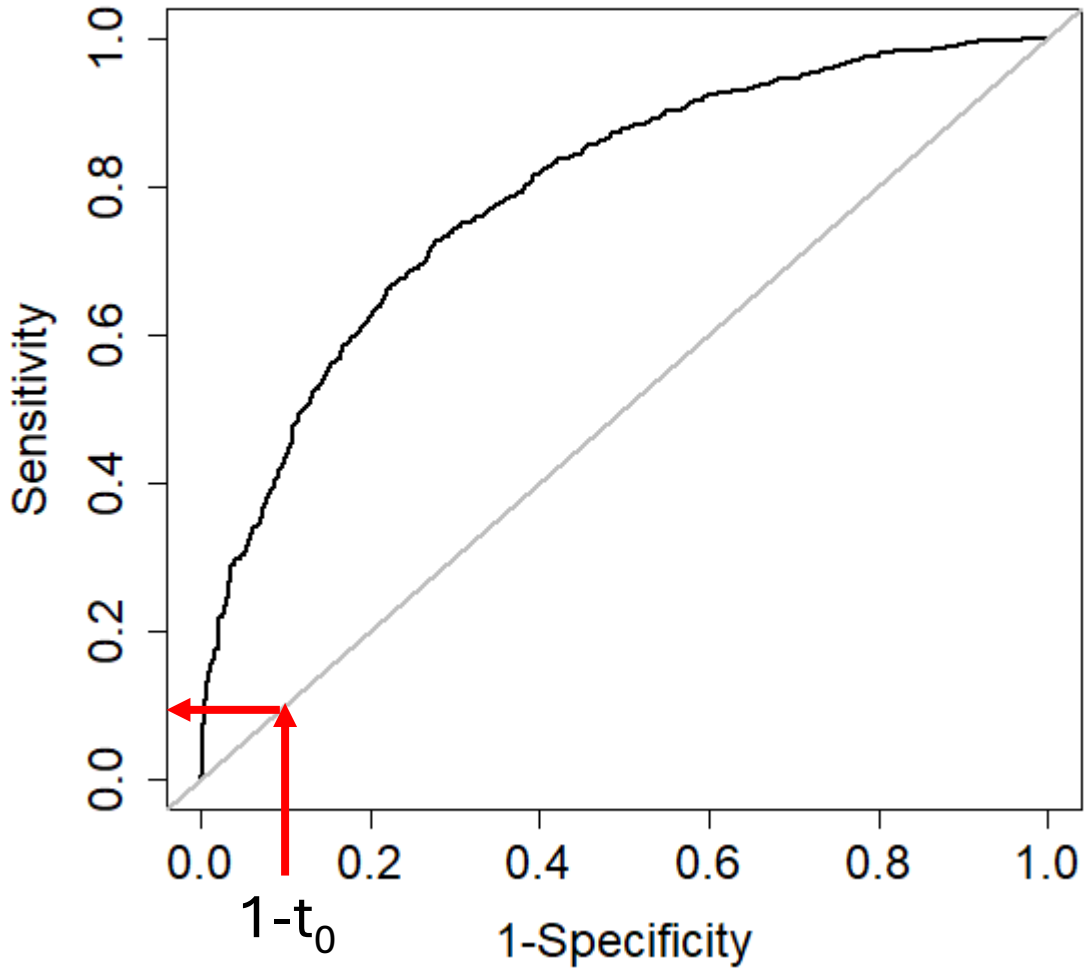
Table 13-5
BMI Measurements

Men	Women
23.8 (11.5)	19.6 (2.5)
23.2 (9)	23.8 (11.5)
24.6 (14)	19.6 (2.5)
26.2 (17)	29.1 (22)
23.5 (10)	25.2 (15.5)
24.5 (13)	21.4 (5)
21.5 (6)	22.0 (7)
31.4 (24)	27.5 (19)
26.4 (18)	33.5 (25)
22.7 (8)	20.6 (4)
27.8 (20)	29.9 (23)
28.1 (21)	17.7 (1)
25.2 (15.5)	
$n_1 = 13$	$n_2 = 12$
$R_1 = 187$	$R_2 = 138$

ROC Curve

Hypothesis Testing

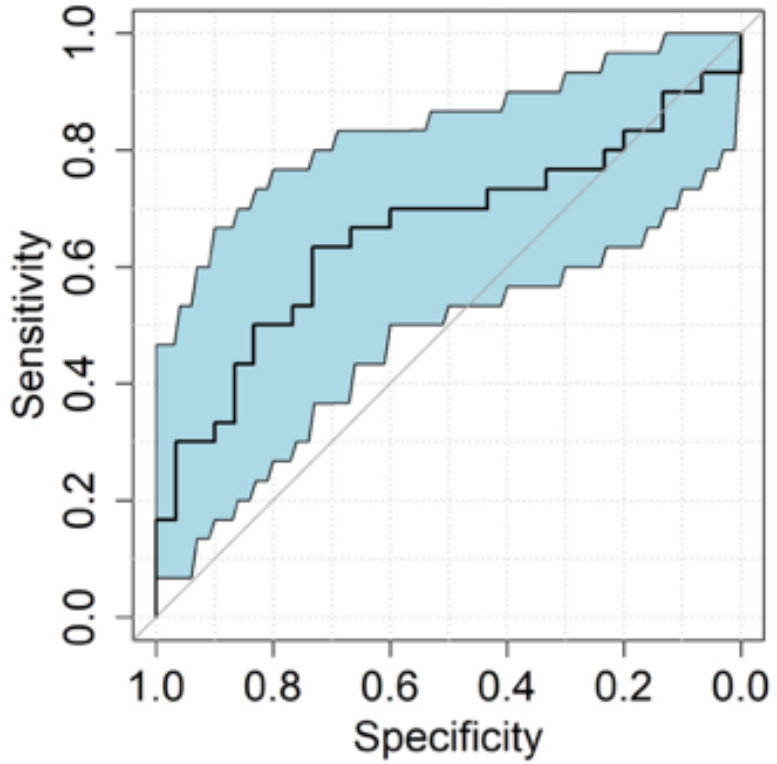
- H_o: AUC=AUC$_0$ vs H_A: not H$_0$
- H_o: PAUC=PAUC$_0$ vs H_A: not H$_0$
- H_o: St at t$_0$ Sp=t$_0$ vs H_A: not H$_0$
- Note that useless biomarker has AUC of 0.5; PAUC= $t_0^2/2$; St at t$_0$ Sp=t$_0$



Hypothesis Testing

- The R pROC package did not provide p-values.
- It only provides confidence intervals.

- Pepe, M.S., 2003. *The statistical evaluation of medical tests for classification and prediction*. Oxford university press.



Chi-square test

- Suppose we will evaluate biomarker at a sensitivity at a fixed specificity.
- Let c be the cutoff corresponding to this sensitivity and specificity.
- For example,

	Case	Control	Total
Value≥c	a	b	a+b
Value<c	c	d	c+d
Total	a+c	b+d	n=a+b+c+d

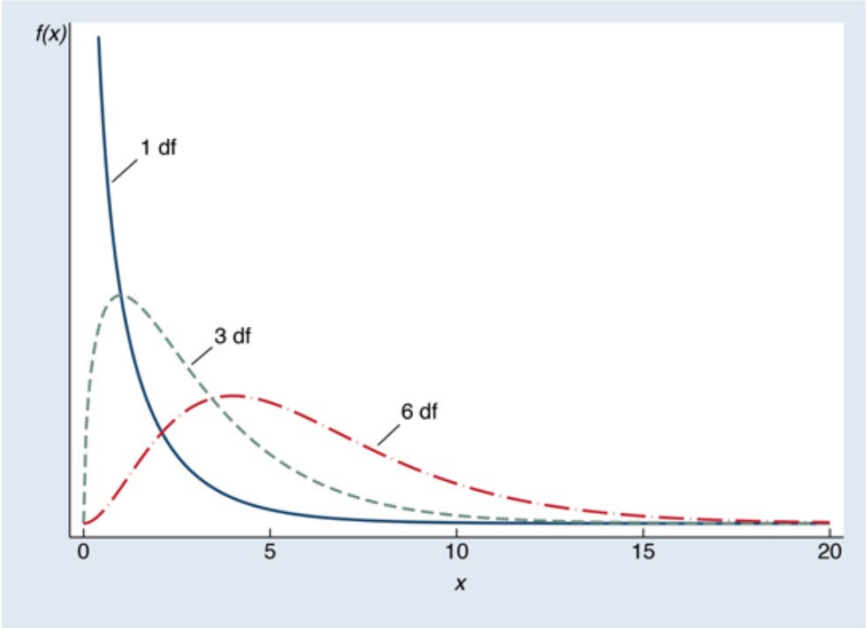
Chi-square test

- H_o: There is no association between biomarker (V1) and outcome (V2) variables vs H_a: Not H_o
- Under the null, P(V1="Case" & V2="Case")=P(V1="Case")P(V2="Case").
    - P(V1="Case")= $\frac{(a+b)}{n}$
    - P(V2="Case")= $\frac{(a+c)}{n}$
    - Expected number of V1="Case" & V2="Case" is $n\frac{(a+b)}{n}\frac{(a+c)}{n} = \frac{(a+b)(a+c)}{n}$, which should be close to "a" if the H_o is true.

- Expected counts:

Variable1	Variable2		Total
	Case	Control	
Case	(a+b)(a+c)/n	(a+b)(b+d)/n	a+b
Control	(c+d)(a+c)/n	(c+d)(b+d)/n	c+d
Total	a+c	b+d	n

- $\chi^2 = \sum_i^2 \sum_{j=1}^2 \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ has a chi-square distribution with df=1
  - O_{ij} : Observed count for the ith row and jth colume
  - E_{ij} : Expected count for the ith row and jth colume
- Compute p-value based on χ^2 statistics.



- Chi-square test requires E_{ij}>5.
- Otherwise, Fisher's exact test can be used.

Multiple testing

The Bonferroni correction

- If we perform null hypotheses multiple times, the risk of making a type 1 error increases.
- Let H_{0j}: jth biomarker is useful, j=1,2,…,M.
- Probability of falsely rejecting at least one null hypothesis is
  - 0.05 if M=1,
  - 0.0975 if M=2,
  - 0.401 if M=10,
  - 0.994 if M=100.
- The Bonferroni correction:
  - Reject H_{0j}, j=1,2,…,M, if its p-value is less than α/n, instead of α.
  - Very conservative. We will hardly reject any of H_{0j}.

False discovery rate

- Suppose we test n hypothesis tests, n_0 of which are true and n-n_0 of which are false.

	H_0 True	H_0 False	Total
Fail to reject H_0	U	T	M-R
Reject H_0	V	S	R
Total	M_0	M_1	M

- V is number of false discovery.
- T is number of false negative.
- Type I error is E(V)/ M_0
- Type II error is E(T)/ M_1
- The power is 1-E(T)/M_1

False discovery rate

- FDR=V/R (or often E(V/R)).
- That is, FDR=E(Number of null hypothesis falsely rejected/number of null hypothesis rejected)
- For NAPPA, FDR=E(Number of biomarkers incorrectly declared significant/number of biomarkers declared significant.)
- This is fraction of discoveries that are false positive.

- By controlling the FDR, we are guaranteeing that we don't have too many false discoveries.
- If we use an FDR threshold of 0.2, then no more than 20% of the null hypotheses that we reject were actually true. (Unfortunately, we don’t know which ones.)

Benjamini-Hochberg Algorithm for FDR Control

- Fix the false discovery rate, α.
- Let p(1) ≤ p(2) ≤ … ≤ p(M) denote the ordered p-values for the M hypothesis tests.
- Find the maximum k such that $\frac{M \times p(k)}{k} \leq \alpha$.
- Reject all tests for which p-values $\leq p(k)$

- Less conservative.

Summary

- Exploratory data analysis is the first step of the data analysis, preceding statistical hypothesis testing and ROC curve analysis.
- P-value based biomarker discovery is generally more preferred.
- Adjusted p-values should be used for NAPPA data analysis.