Accelerating Impact: Immersive Summer Bootcamp in Implementation Science and Biostatistics

Georgian Implementation Science Fogarty Training (GIFT) Program

Ilia State University & Yale University

Yale Center for Analytical Sciences

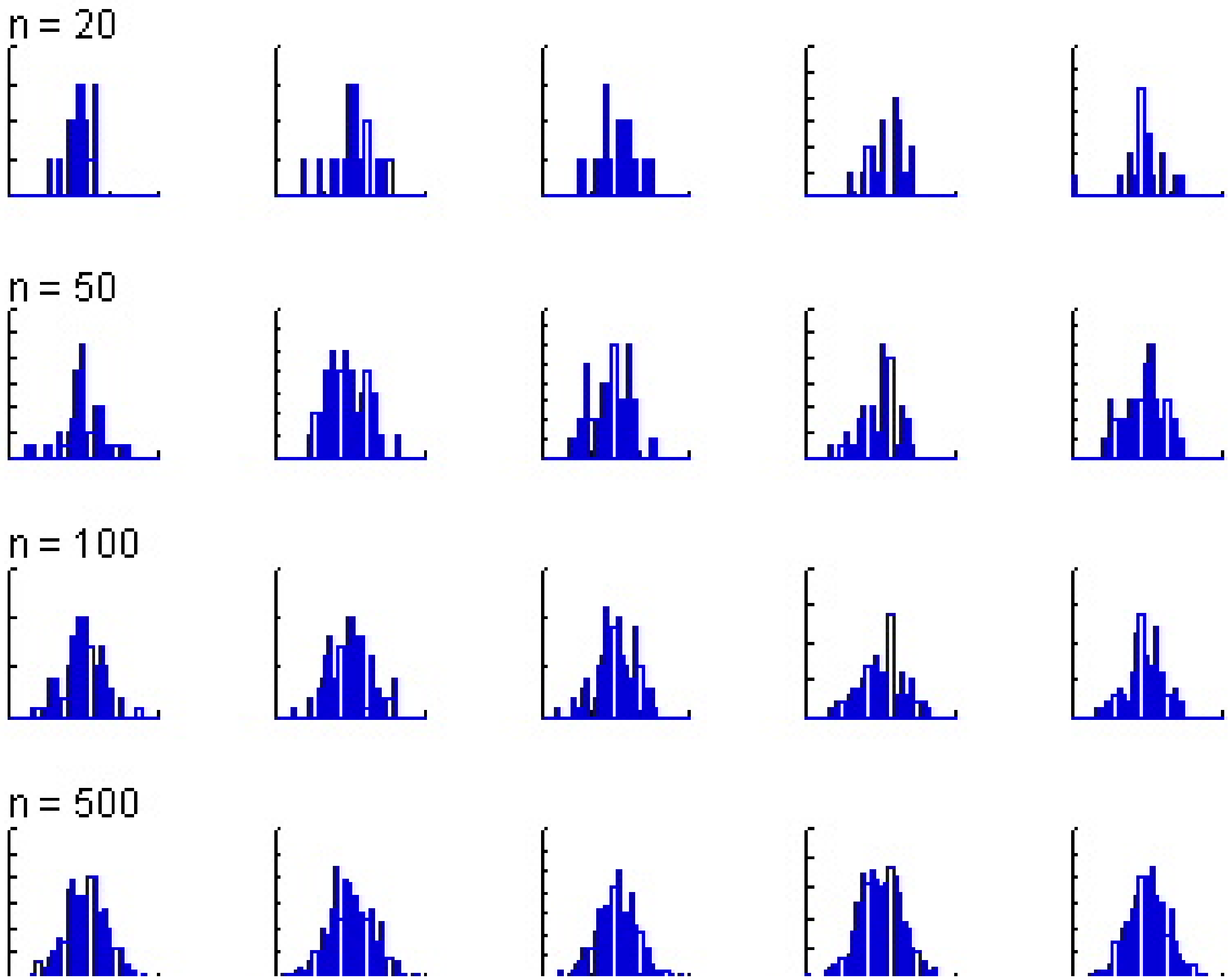
Estimation, Hypothesis Testing, Comparing Means

Parametric

Statistical Inference

- Draw conclusions about a population from a sample
- Two Approaches
  - – Estimation
  - – Hypothesis Testing

Statisticians fall asleep faster by taking a random sample of sheep.



n = 20
n = 50
n = 100
n = 500

Estimation

- Point Estimates – summary statistics from sample to give an estimate of the true population parameter

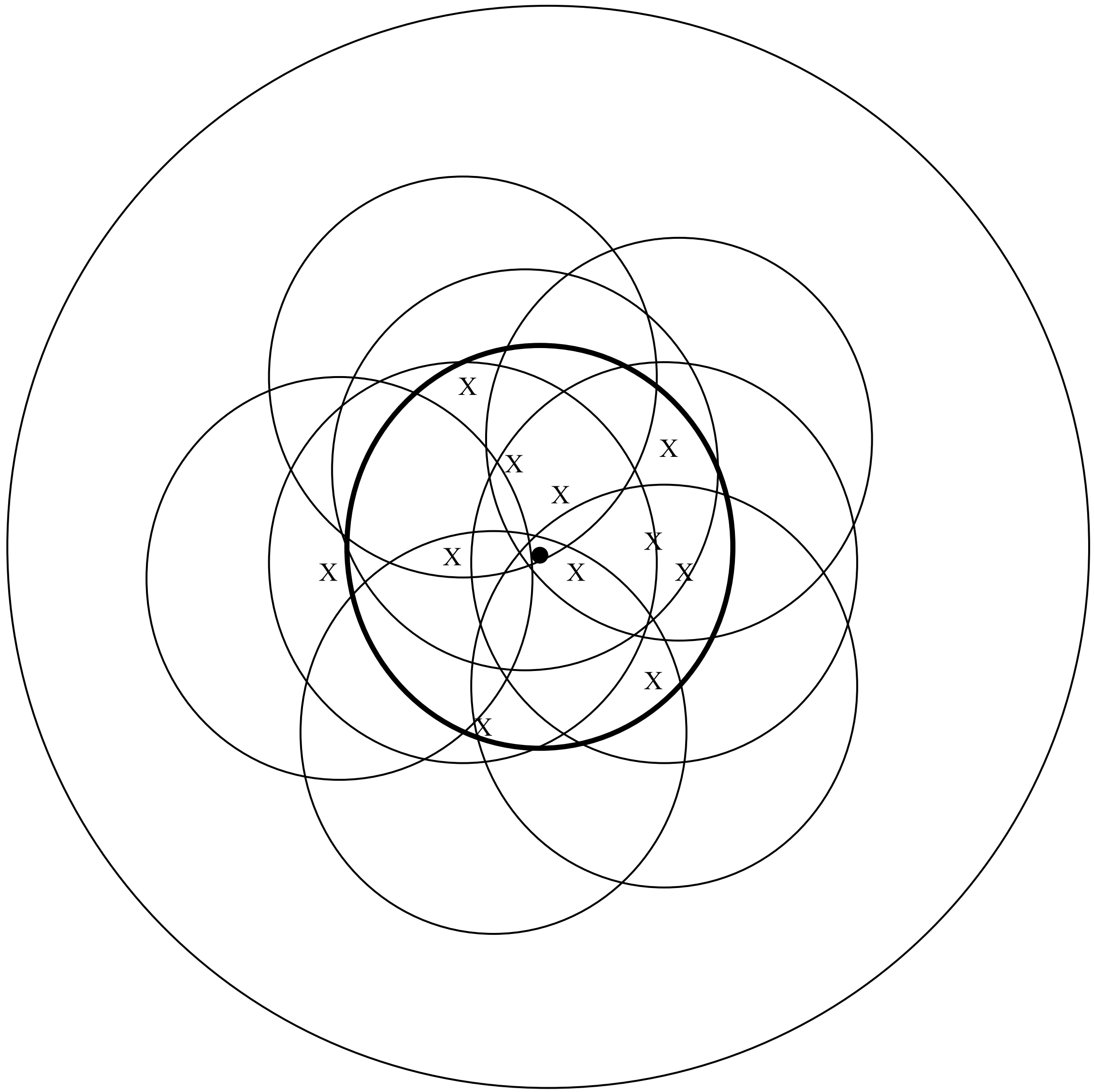
$$\bar{x} \rightarrow \mu$$

$$s \rightarrow \sigma$$

$$p \rightarrow \pi$$

- Confidence Intervals – indicate the variability of point estimates from sample to sample

Confidence Interval



Confidence Interval of Mean

Large Sample ($n \geq 30$)

$$\overline{X} \pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

Where $1-\alpha$ corresponds to the level of confidence or the chance that range of guesses contains the true population mean

Example – 95% Confidence Interval implies that you're 95% sure that the range includes pop. mean

Confidence Interval of a Mean

Large Sample Example

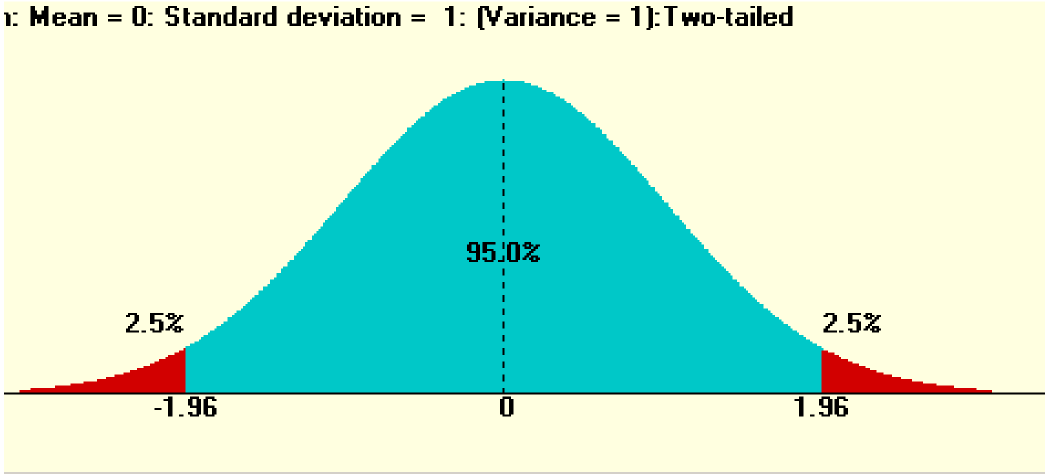
- Measure serum cholesterol in 100 adults and find

$$\bar{x} = 6.7 \text{ mmol/L}$$
$$s = 1.2$$

- Calculate a 95% CI.

Confidence Interval of a Mean
Large Sample Example

- For a 95% CI - $\alpha=0.05$
  - Look up $z_{0.025}$



- $z_{0.025}=1.96$

Confidence Interval of a Mean

Large Sample Example

$$6.7 \pm 1.96 \times \frac{1.2}{\sqrt{100}}$$

$$6.7 \pm 0.24$$

$$6.46, 6.94$$

- Interpretation – if we were to repeatedly take samples of the same size from the population, 95% of our intervals would contain the true population mean, i.e. we're 95% confident that our true mean is covered by this interval

Confidence Interval of a Mean

Small Sample Size (n<30)

- Problem 1: CLT no longer guarantees that x-bar has a normal sampling distribution

- Problem 2: We don't know σ and the sample SD (s) can provide poor approximations

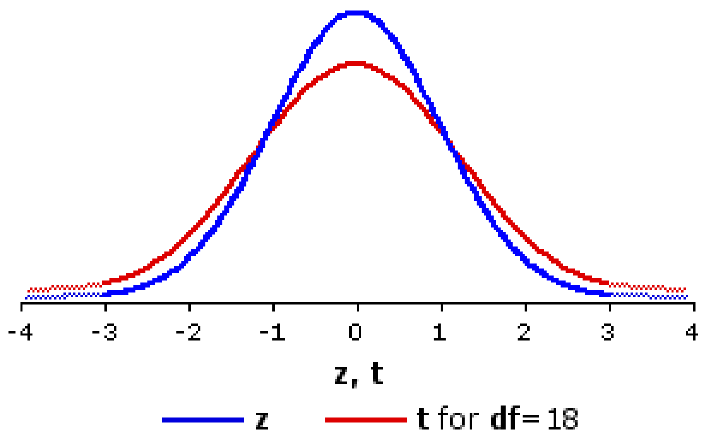
Confidence Interval of a Mean

Small Sample Size (n<30)

- Solve by replacing z with t

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$



Confidence Interval of Mean

Small Sample (n<30)

$$\overline{\mathrm{x}} \pm \mathrm{t}_{\alpha/2,n-1} \times \frac{s}{\sqrt{n}}$$

Suppose only have n=23

$$6.7 \pm 2.074 \times \frac{1.2}{\sqrt{23}}$$

$$6.7 \pm 0.52$$

$$6.18, 7.22$$

" I got the instructions from my Statistics Professor. He was 80% confident that the true location of the restaurant was in this neighborhood."

Degrees of Freedom (d.f.)

- The number of observations allowed to vary

10

20

30

40

<u>25</u>

$\bar{x} = 25$

- there are 5 independent obs that are allowed to vary, but if I know 4 obs and the mean, I must know the value of the 5th obs

Hypothesis Testing

- Like CIs, the purpose is to permit generalizations from sample to population

- The numerical value corresponding to a comparison of interest is called the “effect”
  – We can use hypothesis tests to tell us whether these effects are likely to exist or whether they’re likely to have arisen from random variation

Steps in Hypothesis Tests

Step 1. State the research question in a hypothesis

Null Hypothesis

H_o – no difference – effect is zero

μ=cst

Alternative Hypothesis

H_a – some difference – effect is not zero

Directionality – one-tailed vs two-tailed

$\mu < cst$

$\mu \neq cst$

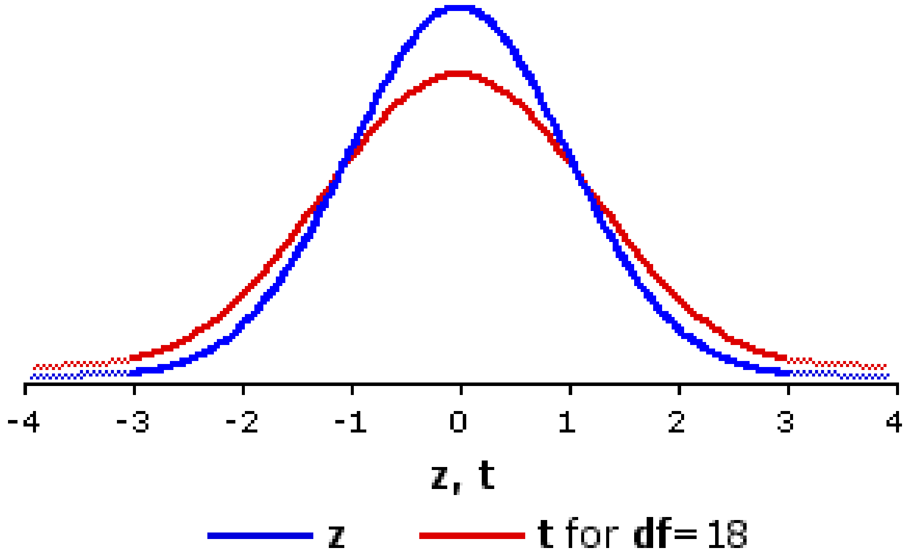
Step 2. Choose appropriate test statistic

A theme for the rest of the course

When comparing 1 or 2 means use t or z

When comparing 1 or 2 proportions and satisfy 5/n criteria use z

- Under the null hypothesis (i.e. assume that there is no difference) – the sampling distribution of the test statistic should follow the corresponding distribution

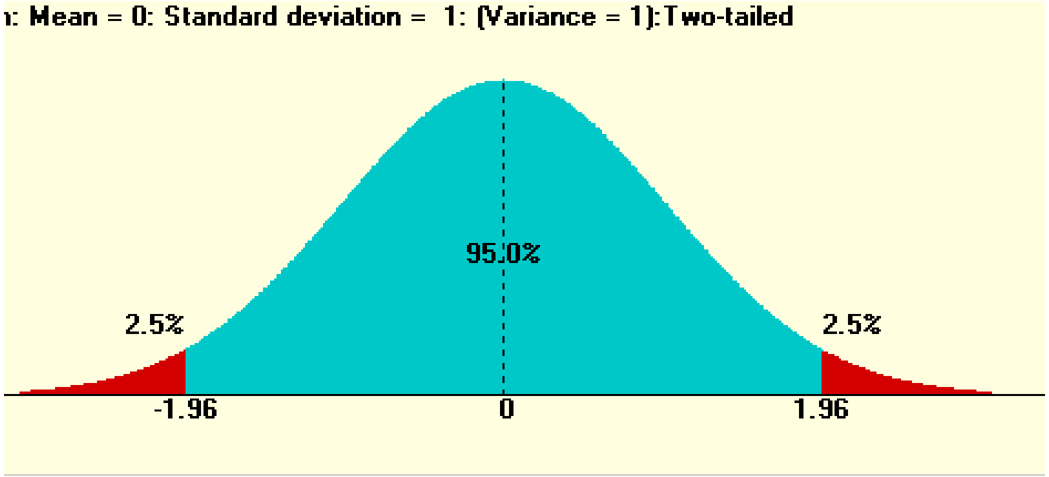


Step 3. Choose the level of significance – α

- how much confidence do you want in decision to reject the null hypothesis
- α is also considered to be the type I error or false positive level (prob. of incorrectly rejecting the null when it is true)
- typically set at 0.05

Step 4. Determine the critical value of the test statistic that must be obtained to reject the null hypothesis

- Example – two-tailed 0.05 significance level for z-test



Step 5. Calculate the test statistic

-Example

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

Step 6. Compare the statistic to the critical value

- If test statistic is more extreme than the critical value then reject the null and accept the alternative
- If test statistic is not more extreme than the critical value then Do Not Reject

DO NOT ACCEPT THE NULL!

HT Example – One Sample Test

- 23 subject sample

$$\bar{x} = 6.7 \quad s = 1.2$$

Step 1. $Ho: \mu = \mu_0 = 5.2$

$Ha: \mu \neq \mu_0 = 5.2$

Step 2. Since <30 subjects use t-statistic

Step 3. Choose a 0.05 significance level

HT Example cont'd

Step 4. Critical value of t

$$23-1=22 \text{ d.f.}$$

$$t_{crit}=\pm 2.074$$

Step 5. Calculate statistic

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} = \frac{6.7 - 5.2}{1.2 / \sqrt{23}} = 5.99$$

HT Example cont'd

Step 6. $5.99 > 2.074$

Reject the null hypothesis

Conclude: The mean serum cholesterol of Population Y is significantly larger than 5.2

Errors in Hypothesis Testing

Truth

Test Conclusion

	Difference Exists (Ha)	**No Difference Exists (Ho)**
Difference Exists (Reject Ho)	1-β	Type I Error = α
No Difference Exists (Do not reject Ho)	Type II Error = β	1-α

<u>Type I Error (false positive)</u> – incorrectly reject the null hypothesis when no difference exists

<u>Type II Error (false negative)</u> – incorrectly say no difference when a difference exists

<u>Power (1-β)</u> – ability to detect a difference when a difference exists

Similarities between CI and Hypothesis Test

- Use the same x-bar, t and standard error

- Confidence interval suggests that 5.2 is an unlikely value for the population Y mean as does the hypothesis test

- Assumptions
  - Sample of randomly drawn independent observations from the pop. of interest
  - If n<30 the observed values arise from a normal pop. distribution

HT of a Single Proportion
Large Sample Example

- A sample of 80 subjects from the population of New Haven provided an estimate for the prevalence of asthma of 16%. Is this higher than a previously established nation-wide prevalence of 14%?
- Remember that the normal distribution approximates the binomial distribution if n and p are large enough
  - Guideline – p and 1-p are greater than 5/n

HT of a Single Proportion

Step 1.

$$H_o : \pi_{NH} = \pi_0 = 0.14$$

$$H_a : \pi_{NH} > \pi_0 = 0.14$$

Step 2. Appropriate test statistic is z (because passes 5/n rule)

Step 3. Level of significance – α=0.025

- one-tailed – will give same as two-tailed 0.05

HT of a Single Proportion

Step 4. Find critical value

- $z_{crit}=1.96$

Step 5. Calculate z-statistic

$$z = \frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} = \frac{0.16 - 0.14}{\sqrt{0.14(0.86)/80}} = 0.52$$

HT of a Single Proportion

Step 6. 0.52<1.96 DO NOT REJECT

Continuity Correction – since the z distribution is continuous and the binomial distribution is discrete need to correct statistic

$$z = \frac{|p - \pi_0| - (1/2n)}{\sqrt{\pi_0(1-\pi_0)/n}} = 0.35$$

CI for a Single Proportion

$$100(1-\alpha)\%\ \mathrm{CI} = p \pm z_{\alpha/2} \times se(p)$$

$$\text{where } se(p) = \sqrt{\frac{p(1-p)}{n}}$$

$$99\%CI = 0.16 \pm 2.575 \times \sqrt{\frac{0.16(1-0.16)}{80}}$$

0.05,0.27

95% CI = 0.08, 0.24

MAYBE USE AN EXAMPLE HERE IN R COMPARING TWO MEANS AND TWO PROPORTIONS

Analysis of Variance (ANOVA)

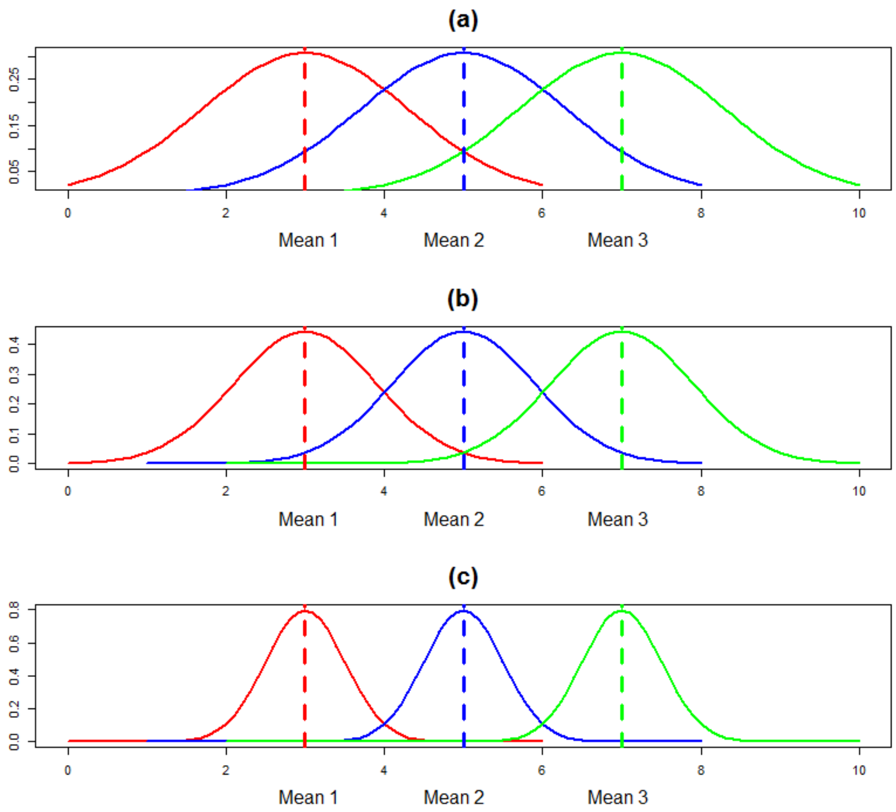
- Used to compare >2 **means**

- Definitions

  - Response variable (dependent) – the outcome of interest – must be continuous

  - Factors (independent) – variables by which the groups are formed and whose effect on response is of interest – must be categorical

ANOVA



- Means are same distance apart in all graphs
- Within-group variation is biggest for (a) and smallest for (c)
- How do we quantify that?

ANOVA

- Quantify:
  - Between-Group Variation: how far group means are from the grand mean

vs.

  - Within-Group Variation: how far individual observations are from their group mean

ANOVA
Assumptions

- Normality
- Homogeneity of variance
- Random sample of independent observations

Multiple Comparisons

- After rejecting null hypothesis of ANOVA, most often we'd like to know which means differ from another
  - ~~Solution: Use individual t-tests to compare all pairs~~
- A 0.05 significance level – there's a 5% chance of a false positive
- The more tests we conduct the greater chance of a false positive

$$1-(1-\alpha)^n$$

A priori Comparisons or Contrasts

- Planned before analysis takes place
- In this case you actually don't need the ANOVA (but in practice it's often done)
- Bonferroni method – conservative but simple
  - Divide the level of significance by the number of comparisons to be made
    - Example: 3 comparisons $\frac{0.05}{3} = 0.017$

A priori Comparisons

- Dunnet – when wish to make all comparisons compared to one gr
  - Example: compare different doses
    placebo

"It didn't cure their allergies, but the treatment group did have 18% fewer cavities than the group taking the sugar pill."

Post-hoc Comparisons

- After ANOVA has resulted in a significant F-test
  - Tukey – can perform all pairwise comparisons
  - Scheffe – most versatile – controls for all possible linear contrasts – most conservative

$$Ho: \frac{\mu_1 + \mu_2}{2} = \frac{\mu_3 + \mu_4}{2}$$

MAYBE USE AN EXAMPLE HERE IN R ONE-WAY ANOVA

Factorial ANOVA (Two-Way)

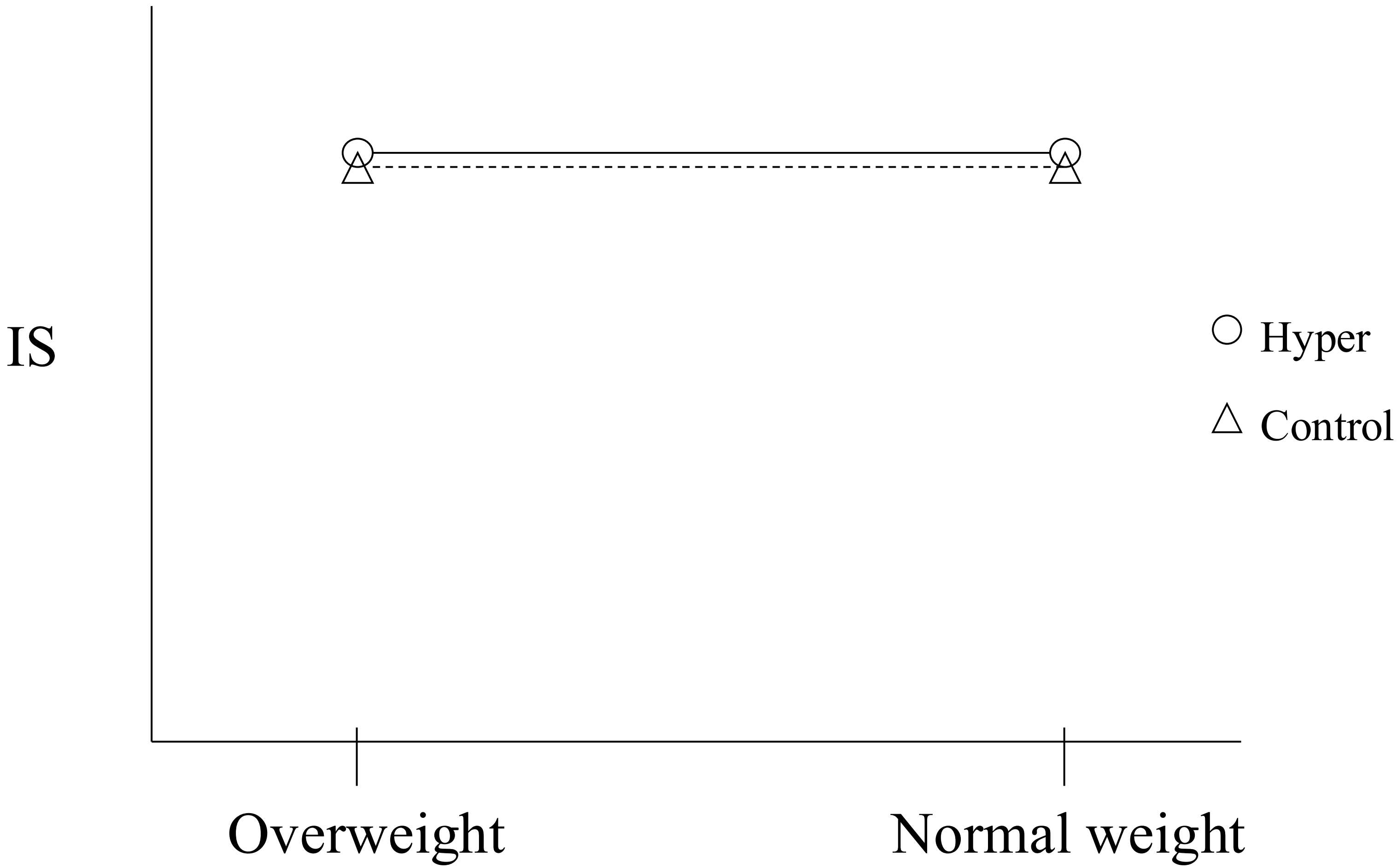
- ANOVA with 2 or more factors
- Example: Is insulin sensitivity (IS) dependent on thyroid level and/or body mass index?

		Overweight	
		Yes	No
Hyperthyroidism	Yes	XXXXXXXXXXXX	XXXXXXXXXXXX
	No	XXXXXXXXXXXX	XXXXXXXXXXXX

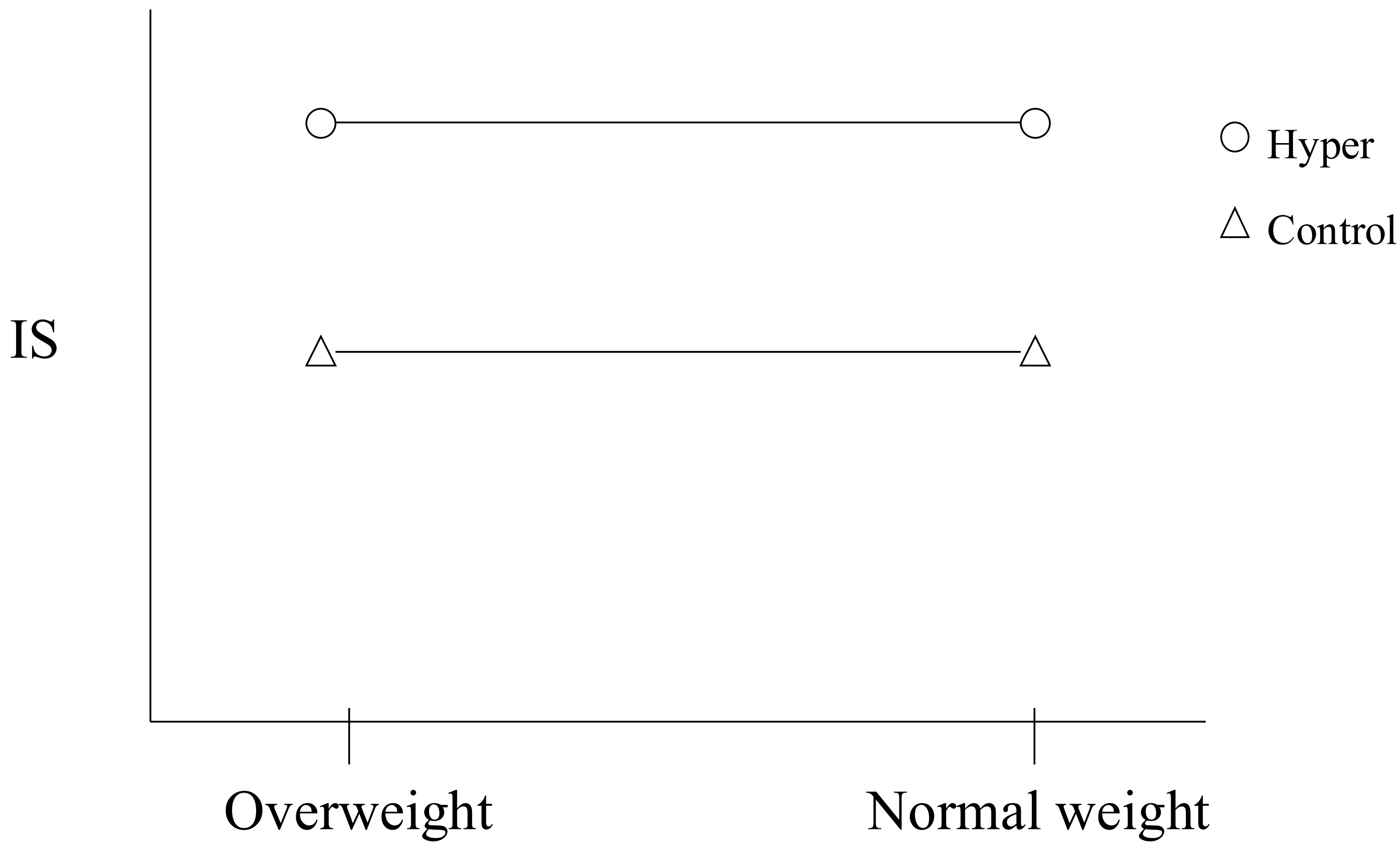
Factorial ANOVA Example

- Three questions to consider
  - Does IS differ b/w those with and those without hyperthyroidism?
  - Does IS differ b/w overweight and normal weight subjects?
  - Is there an additional effect of being both overweight and hyperthyroidal?

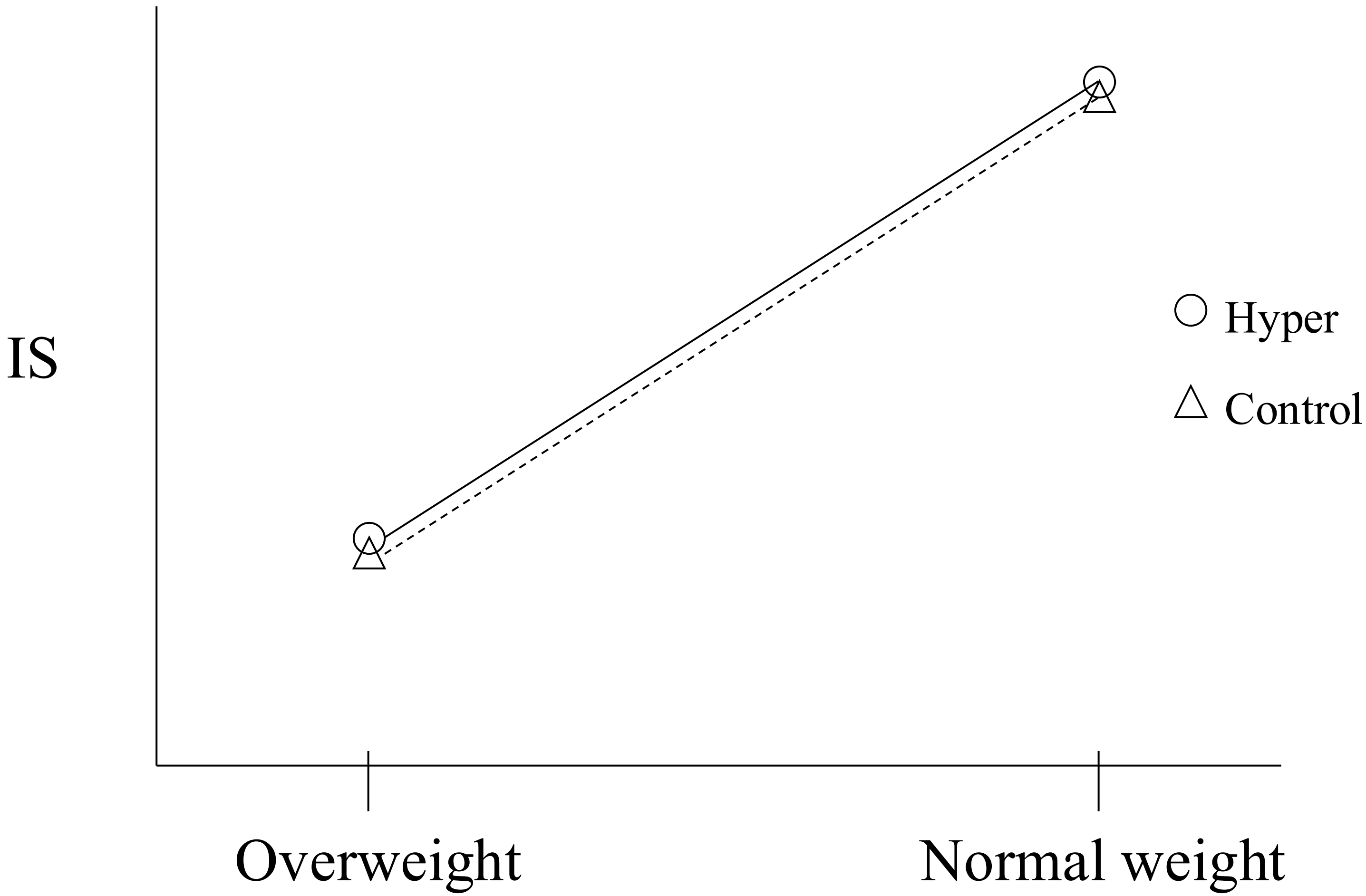
No Effects



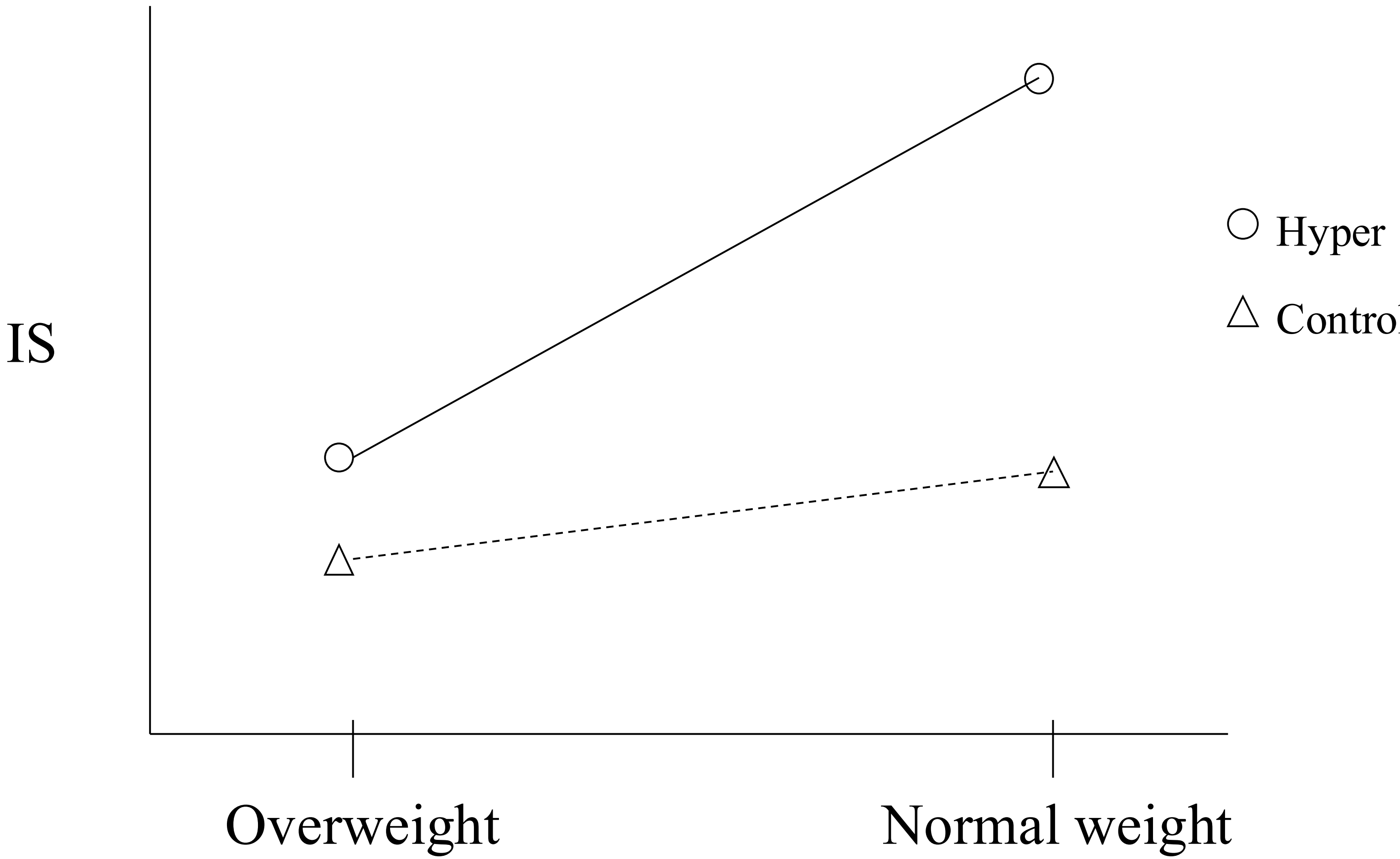
Main Effect of Thyroid Level



Main Effect of Overweight



Interaction of Thyroid Level and Overweight

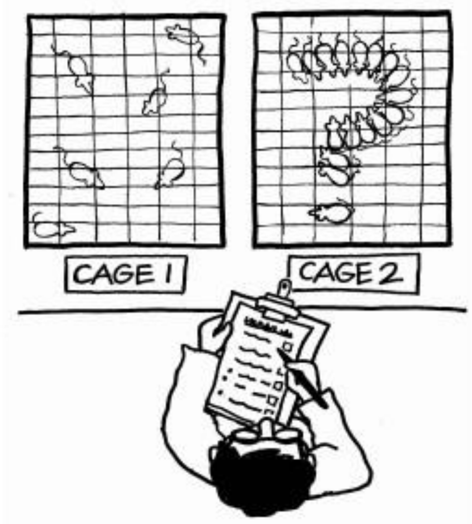


Factorial ANOVA: Sources of Variation

- Sum of Squares for Factor 1 – Thyroid level
- Sum of Squares for Factor 2 – Overweight
- Sum of Squares for Interaction
- Sum of Squares for Error

Assumptions

- Normality
- Homogeneity of variance
- Independence



Despite the clear difference between the treated and control groups, something made him question the data.

MAYBE USE AN EXAMPLE HERE IN R TWO-WAY ANOVA

Estimation, Hypothesis Testing, Comparing Means

Non-parametric

Parametric Tests

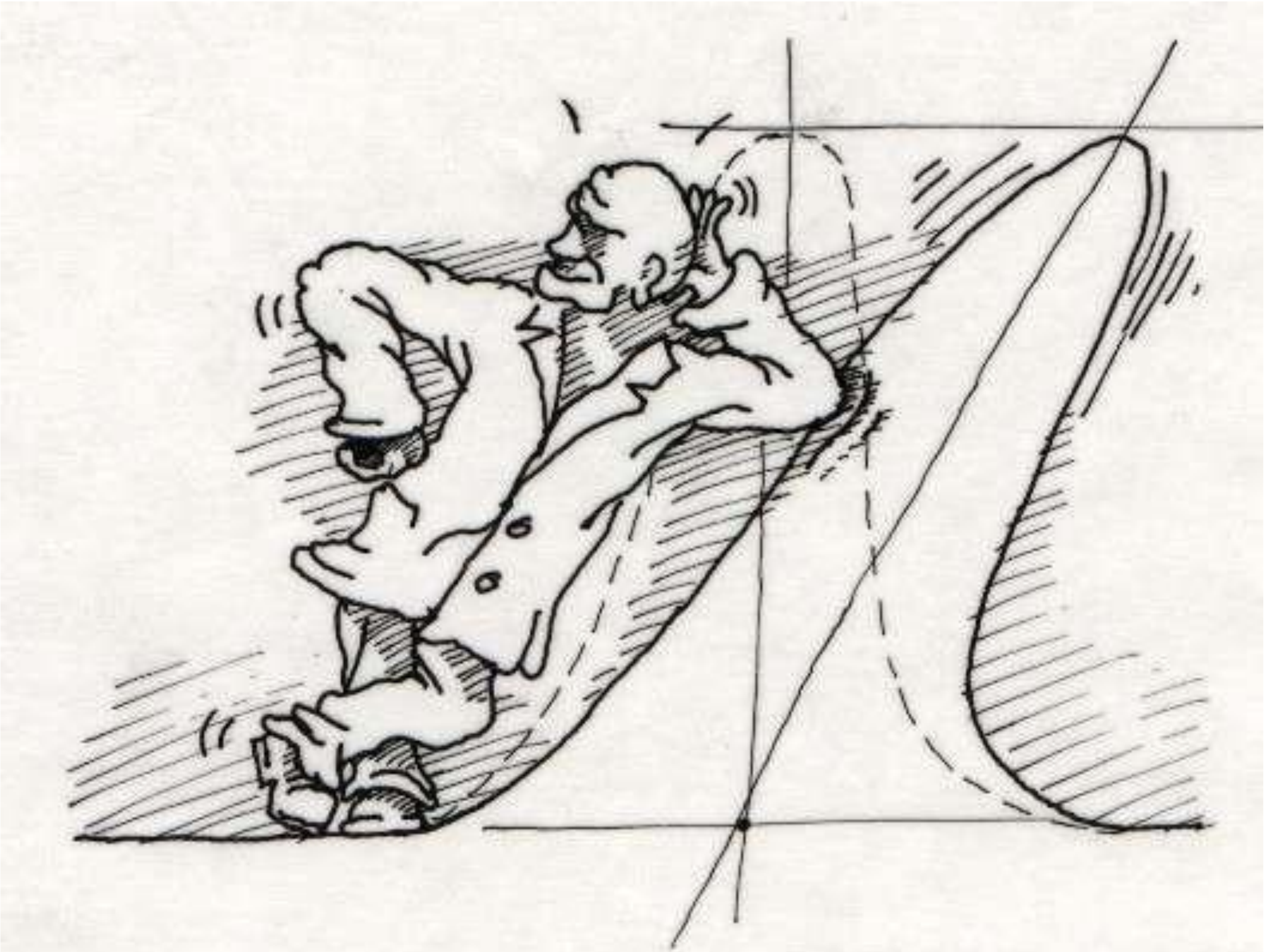
- Normality – the sampling distribution will follow a z or t-distribution provided that:
  - Population values are normally distributed
  - Sample size is sufficiently large (CLT; $n>30$)

- Equal Variances – not a problem if group sizes are similar

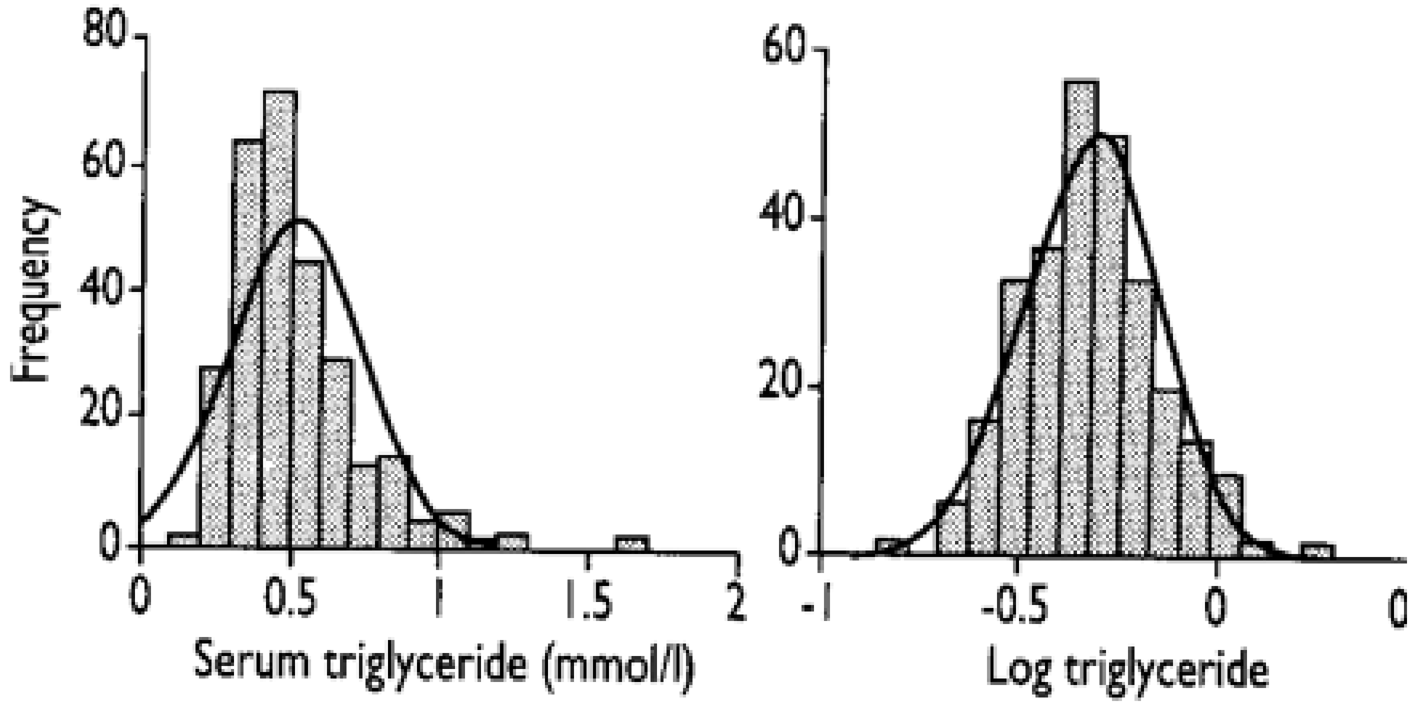
- Scale of measurement is interval – equally spaced

Parametric Statistics

- In parametric statistics we make assumptions about the population that produced the data (i.e. normality, equal variances) to inform us about the sampling distribution of a certain statistic. We then make inferences based on the behavior of sampling distribution.



Positively Skewed Variable
Log Transformation



Non-parametric tests

- Designed for the researcher who knows nothing about the parameters of a variable of interest
  - Distribution-free methods – make no assumptions about the distribution of variable in the population

When Do We Use Non-parametric?

- Definite
  - Variable not normally distributed in population and transformation doesn’t normalize
  - Outcome is ordinal
  - Some values are off the scale
- Maybe
  - Not sure outcome is normally distributed, but not sure that it’s not

Non-parametric vs Parametric

	Normal	**Non-normal**
Large Sample	Ok to use nonparametric (less powerful)	Ok to use parametric - CLT
Small Sample	Nonparametric will lack power (p-values too high)	Parametric may be invalid (p-value is inaccurate)

Inference for a Single Population
Sign Test

- Test comparing a population median to a constant
- Example
  - Under current treatments, the median time to cancer remission is 4.5 years. A new treatment in 7 patients produces the following remission times:
    - 5.3 7.3 3.6 5.2 6.1 4.8 8.4

Sign Test

One - tailed	Two - tailed
$H_0 : \eta = 4.5$	$H_0 : \eta = 4.5$
$H_a : \eta > 4.5$	$H_a : \eta \neq 4.5$
$(H_a : \eta < 4.5)$	

- If the null hypothesis is true, then expect to have the same number of observations above and below this constant in the sample

  - Test Statistic –

    S_1 - # of sample measures $> \eta_0$ Larger of S_1 and S_2

    (S_2 - # of sample measures $< \eta_0$)

Sign Test

- Determine the observed significance level

<u>One - tailed</u>	<u>Two - tailed</u>
$p(x \geq S)$	$2p(x \geq S)$

Where x has a binomial distributi on with parameters :

$\text{n} = \#\text{ of subjects} \qquad \text{p} = 0.50$

(Table II Binomial Probabilit ies)

- Reject if p≤α

Sign Test

- Example

$$S_1 = 6 \quad S_2 = 1 \quad \rightarrow \quad 2p(x \geq 6) = 2(1 - p(x \leq 5))$$

$$\text{for n} = 7 \quad \rightarrow \quad 2(1 - .937) = 0.126$$

Do Not Reject

- For samples n≥10 can use the normal approximation

$$z = \frac{(S - 0.5) - 0.5n}{0.5\sqrt{n}}$$

Inference for 2 Independent Populations
Wilcoxon Rank Sum Test

- Researcher tests the hypothesis that hearing impaired children have greater visual acuity compared to children without impairment by measuring eye movement rates in 10 deaf and 10 hearing children

One-tailed	Two-tailed
$H_0: D_1 = D_2$	$H_0: D_1 = D_2$
$H_a: D_1$ is shifted to right	$H_a: D_1$ is shifted left or right
[D_1 is shifted to left]	

Wilcoxon Rank Sum Test

Deaf	Rank	Hearing	Rank
2.75	18	1.15	1
3.14	19	1.65	6
3.23	20	1.43	4
2.30	15	1.83	8.5
2.64	17	1.75	7
1.95	10	1.23	2
2.17	13	2.03	12
2.45	16	1.64	5
1.83	8.5	1.96	11
2.23	14	1.37	3

Rank Sums $T_1 =$ 150.5 $T_2 =$ 59.5

Wilcoxon Rank Sum Test

- Test statistic – the rank sum (T_1 or T_2) of the smaller sample (if n_1=n_2, either will do)
- Rejection Region

<u>One - tailed</u>

$T_1 \geq T_U \ [\text{or } T_1 \leq T_L] \ \text{ if } \ n_1 < n_2$

$T_2 \leq T_L \ [\text{or } T_2 \geq T_U] \ \text{if } \ n_2 < n_1$

<u>Two - tailed</u>

$T \leq T_L \text{ or } T \geq T_U$

Where T_L and T_U are from critical values from the table for Wilcox on Rank Sum boundaries

Wilcoxon Rank Sum Test

- Example

$$H_0 : D_D = D_H$$
$$H_a : D_D \text{ is shifted to right of } D_H$$

- Since $n_1=n_2$ can use T_1 or T_2
  - $T_L=83$ $T_U=127$
    - 150.5>127 OR 59.5<83
    - Reject Null and conclude the distribution of eye movement rate is shifted to the right for hearing impaired children

Wilcoxon Rank Sum Test
Large Sample Approximation

- If n_1 and $n_2 \geq 10$ can use normal approximation

$$z = \frac{T_1 - \frac{n_1(n_1 + n_2 + 1)}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

$$= \frac{150.5 - \frac{10(10 + 10 + 1)}{2}}{\sqrt{\frac{10 \times 10(10 + 10 + 1)}{12}}} = 3.44 \quad > 1.645 \text{ Reject Null}$$

Inference for 2 Dependent Populations
Wilcoxon Signed Rank Test

- Researcher wishes to compare ratings on a pain scale in subjects before and after treatment

<u>One - tailed</u>

$H_0 : D_{PRE} = D_{POST}$

$H_a : D_{PRE}$ is shifted to right

$[D_{PRE}$ is shifted to left$]$

<u>Two - tailed</u>

$H_0 : D_{PRE} = D_{POST}$

$H_a : D_{PRE}$ is shifted left or right

Wilcoxon Signed Rank Test

Subject	Pre	Post	Diff	|Diff|	Rank	
1	6	4	2	2	5	+
2	8	5	3	3	7.5	+
3	4	5	-1	1	2	-
4	9	8	1	1	2	+
5	4	1	3	3	7.5	+
6	7	9	-2	2	5	-
7	6	2	4	4	9	+
8	5	3	2	2	5	+
9	6	7	-1	1	2	-
10	8	2	6	6	10	+
~~11~~	~~5~~	~~5~~	~~0~~	~~0~~		

T_+ = sum of positive ranks = 46

T_- = sum of negative ranks = 9

Wilcoxon Signed Rank Test

- Test Statistic

One - tailed	Two - tailed
T_- $[T_+]$	the smaller of T_- or T_+

- Rejection region

One - tailed	Two - tailed
$T_- \leq T_0$ $[T_+ \leq T_0]$	$T \leq T_0$

Where T_0 is the critical value from Table for Wilcoxon Signed Rank

Inference for >2 Independent Populations

Kruskal Wallis H-test

- A hospital administrator would like to compare the number of unoccupied beds in 3 hospitals
  - Randomly selects 10 different days for each hospital and records number of unoccupied beds

$$H_0: \text{the probability distributions are equal in location}$$

$$H_a: \text{At least one distribution differs in location}$$

Kruskal Wallis H-test

Hospital 1	Rank	Hospital 2	Rank	Hospital 3	Rank
6	5	34	25	13	9.5
38	27	28	19	35	26
3	2	42	30	19	15
17	13	13	9.5	4	3
11	8	40	29	29	20
30	21	31	22	0	1
15	11	9	7	7	6
16	12	32	23	33	24
25	17	39	28	18	14
5	4	27	18	24	16

$R_1 =$ 120 $R_2 =$ 210.5 $R_3 =$ 134.5

Kruskal Wallis H-test

- Test Statistic

$$\mathrm{H} = \frac{12}{n(n+1)} \sum \frac{R_j^2}{n_j} - 3(n+1)$$

- Rejection Region

$$\mathrm{H} > \chi^2_\alpha \text{ with } \mathrm{p-1\, d.f.}$$

Kruskal Wallis H-test

- Example

$$H = \frac{12}{30(31)}\left(\frac{120^2}{10} + \frac{210.5^2}{10} + \frac{134.5^2}{10}\right) - 3(31)$$

$$= 6.097$$

$$\chi^2_{0.05,2} = 5.99$$

6.097>5.99 so reject null and conclude that at least one distribution is different

Kruskal Wallis H-test

- We can equivalently write the H statistic as:

$$\mathrm{H} = \frac{12}{n(n+1)} \sum (\overline{R}_j - \overline{R})^2$$

Mean Treatment Rank

Overall Mean Rank

- What do you do after reject null for H-test?

Randomized Block Design

Friedman F_r Test

- Investigator wishes to compare the time to pain relief following three different medications
  - Each subject receives each medication in random order

$$H_0: \text{the probability distributions are equal in location}$$

$$H_a: \text{At least one distribution differs in location}$$

Friedman F_r Test

Subject	Drug A	Rank	Drug B	Rank	Drug C	Rank
1	9	1	11	2	18	3
2	13	2	13	2	13	2
3	11	1	12	2.5	12	2.5
4	10	1	15	2	16	3
5	9	2	8	1	10	3

$R_1 =$ 7 $\quad R_2 =$ 9.5 $\quad R_3 =$ 13.5

Friedman F_r Test

- Test Statistic

Rank sum for the j^{th} treatment

$$F_r = \frac{12}{bp(p+1)} \sum R_j^2 - 3b(p+1)$$

\# of blocks

\# of treatments

- Rejection Region

$$F > \chi_\alpha^2 \text{ with } p-1 \text{ d.f.}$$

Friedman F_r Test

- Example

$$F_r = \frac{12}{5(3)(4)}(7^2 + 9.5^2 + 13.5^2) - 3(5)(4)$$

$$= 4.3$$

$$\chi^2_{0.05,2} = 5.99$$

4.3 is not greater than 5.99 so do not reject null

EXAMPLES OF NON-PARAMETRIC TESTS

Advantages of Non-parametric Tests

- Make less stringent demands on the data (not as many assumptions).
- Not computationally intensive
- Do not require a reliable underlying measurement scale
- In many cases are not too inefficient compared to parametric techniques

Disadvantages of Non-parametric Tests

- No parameters to describe – can be difficult to make quantitative statements about the populations
- Throw away information