Statistical Tools for Literature and Language Studies

Joel Reyes Noche

jrnoche@adnu.edu.ph
University Research Council
Ateneo de Naga University

March 22, 2010

Measurement

[SC88, p. 32]

Measurement is the process of mapping or assigning numbers to objects or observations. The kind of measurement achieved is a function of the rules under which the numbers are assigned to objects. The operations and relations employed in obtaining the scores define and limit the manipulations and operations which are permissible in handling the scores; the manipulations and operations must be those of the numerical structure to which the particular measurement is isomorphic.

Examples of Measurements

- Communication strategy (approximation, word coinage, circumlocution, literal translation, language switch, appeal for assistance, mime, topic avoidance, message abandonment, repetition, restructuring) [GRO07]
- Oral language proficiency—vocabulary (extremely limited vocabulary, small vocabulary, vocabulary of moderate size, large vocabulary, extensive vocabulary) [GRO07]
- Birth date (month, day, year)
- Age (number of months)

Scales of Measurement

[SC88]

	Measurement scale			
Defining relations	Nominal	Ordinal	Interval	Ratio
Equivalence	✓	✓	✓	✓
Greater than		✓	✓	✓
Known ratio of any two intervals			✓	✓
Known ratio of any two scale values				✓

More Examples of Measurements

What measurement scale is used in each of the following variables?

- Faculty level (assistant instructor, instructor, assistant professor, associate professor, professor)
- Comprehensible input strategy employed by teachers (pronunciation, vocabulary, syntax, discourse) [PGRO06]
- Response to a five-level Likert item (strongly disagree, disagree, neither agree nor disagree, agree, strongly agree)
- Number of questions correctly answered in a 100-item objective-type test [Cor09]

The behavioral and the social sciences commonly use qualitative (nominal and ordinal) data. The chemical, the biological, the physical, and the engineering sciences commonly use quantitative (interval and ratio) data. The statistical tools usually taught in general education undergraduate classes are mostly for quantitative data.

How Probability and Statistics Differ

In *probability*, we are given a (known) population and we make an educated 'guess' about the (unknown) sample.

- If most students in a university are male, then how likely is it that most students in a certain class in the university are male?
- If a coin is fair, then how likely is it that in ten coin flips, exactly five heads turn up?

In *statistics*, we are given a (known) sample and we make an educated 'guess' about the (unknown) population.

- If most students in a certain class in a university are male, then how likely is it that most students in the university are male?
- If exactly five heads turn up in ten coin flips, then how likely is it that the coin is fair?

How Probability and Statistics Are Alike

A *parameter* is a numerical descriptive measure of a population.

A *sample statistic* is a numerical descriptive measure of a sample.

A (stochastic) *experiment* is an act or process of making an *observation* (or *measurement*) that leads to a single *outcome* that cannot be predicted with certainty.

The *probability* of an *event* (a specific set of outcomes) is a number that indicates the likelihood that the event will occur when the experiment is performed.

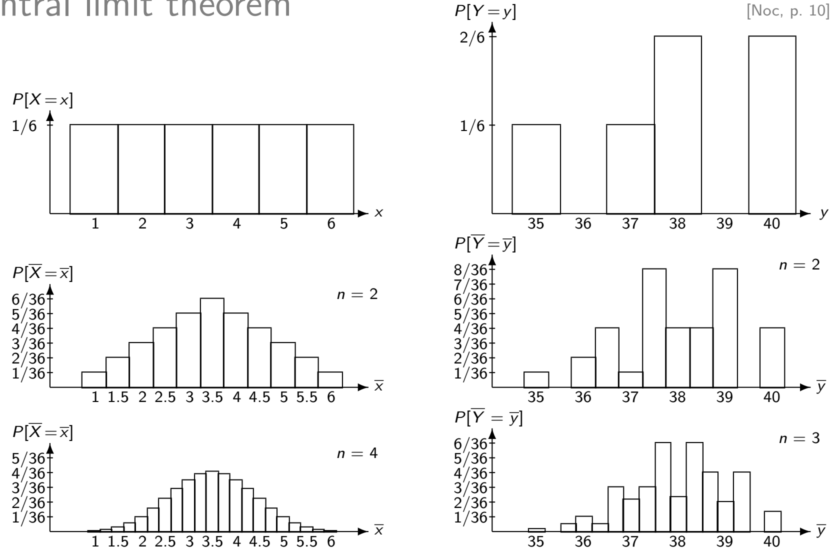
A *random variable* is an assignment of exactly one number to each outcome of an experiment.

A sample statistic is a random variable.

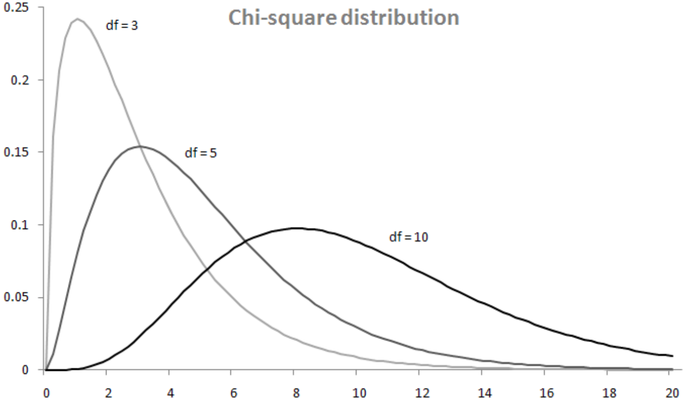
(Probability and random variables are further discussed in the handouts [Noc, pp. 3–10].)

Central limit theorem

[Noc, p. 10]



Degrees of freedom



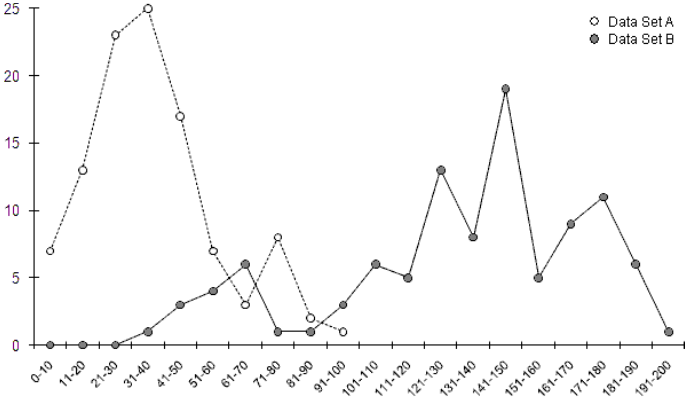
Descriptive Statistics

[MD92]

Descriptive statistics use numerical and graphical methods to look for patterns, to summarize, and to present the information in a set of data.

- *Central tendency* is the tendency of a data set to cluster about certain numerical values.
- *Variability* is the spread of a data set.
- *Skewness* is the nonsymmetry of a data set.
- *Relative standing* is the relative quantitative location of a particular measurement within a data set.

(Descriptive statistics are further discussed in the handouts [Noc, pp. 1–2].)



A: lower central tendency, lower variability, positive skewness
B: higher central tendency, higher variability, negative skewness
The datum 90 is relatively high in *A*, and relatively low in *B*.

Microsoft Excel screenshot and screenshot showing formulas

	A	B	C
1		A	B
2	mean	37.11792	131.6667
3	median	35.75	142
4	standard deviation of a population	19.60441	39.06187
5	standard deviation of a sample	19.69754	39.25477
6	percentile rank	0.997	0.152

	A	B	C
1		A	B
2	mean	=AVERAGE(E2:E107)	=AVERAGE(F2:F103)
3	median	=MEDIAN(E2:E107)	=MEDIAN(F2:F103)
4	standard deviation of a population	=STDEVP(E2:E107)	=STDEVP(F2:F103)
5	standard deviation of a sample	=STDEV(E2:E107)	=STDEV(F2:F103)
6	percentile rank	=PERCENTRANK(E2:E107,90)	=PERCENTRANK(F2:F103,90)

where the data for set A are in cells E2 to E107 and the data for set B are in cells F2 to F103

Inferential Statistics

[MD92]

Inferential statistics (such as estimators and tests of hypotheses) use sample data to make estimates, decisions, predictions, or other generalizations about a larger set of data.

Elements of an Inferential Statistical Problem

1. The population of interest
2. One or more variables (characteristics of the population units) that are to be investigated
3. The sample of population units
4. The inference about the population based on information contained in the sample
5. A measure of reliability for the inference

Estimation

[MD92]

A *point estimator* is a rule or formula that tells us how to use sample data to calculate a single number that can be used as an estimate of a population parameter.

(An example of a point estimator is discussed in the handouts [Noc, p. 9].)

An *interval estimator* is a formula that tells us how to use sample data to calculate an interval that estimates a population parameter.

The *confidence coefficient* is the probability that an interval estimator encloses the population parameter.

The *confidence level* is the confidence coefficient expressed as a percentage.

(Our focus is not on estimation but on tests of hypotheses.)

Elements of a Test of Hypothesis

[MD92]

1. *Null hypothesis* (H_0)
   It is a theory about the values of one or more population parameters. The theory generally represents the status quo, which we accept until proven false.
2. *Alternative (research) hypothesis* (H_a)
   It is a theory that contradicts the null hypothesis. The theory generally represents that which we will accept only when sufficient evidence exists to establish its truth.
3. *Test statistic*
   It is a sample statistic (a number calculated from the sample data) used to decide whether or not to reject H_0.

Elements of a Test of Hypothesis [MD92]

(continued)

4. *Rejection region*
   It is the set of numerical values of the test statistic for which H_0 will be rejected. We choose it so that the probability is α that it will contain the test statistic when H_0 is true. The value of α (the test's *level of significance*) is usually chosen to be small (e.g., 0.01, 0.05, or 0.10).
5. *Assumptions*
   We clearly state any assumptions made about the population(s) being sampled.
6. *Experiment and calculation of test statistic*
   Perform the sampling experiment, and determine the numerical value of the test statistic.

Elements of a Test of Hypothesis [MD92]

(continued)

7. *Conclusion*
   - If the value of the test statistic falls in the rejection region, then we reject H_0 and conclude that H_a is true. The hypothesis-testing process will lead to this conclusion incorrectly only $100\alpha\%$ of the time when H_0 is true.
   - If the value of the test statistic does not fall in the rejection region, then we reserve judgment about which hypothesis is true. We do not conclude that H_0 is true because we do not (in general) know the probability β that our test procedure will lead to an incorrect acceptance of H_0.

An Analogy: The Jury Trial of an Accused Murderer [MD92]

The elements of a test of hypothesis apply to the American jury system of deciding the guilt or innocence of an accused murderer.

1. H_0: The accused is innocent. This is assumed to be true until proven otherwise.
2. H_a: The accused is guilty. This is accepted only when sufficient evidence exists to establish its truth.
3. *Test statistic*: Let x be the number of the jury members who vote "guilty."
4. *Rejection region*: In a murder trial the jury vote must be unanimous in favor of guilt before H_0 is rejected in favor of H_a. For a 12-member jury trial, the rejection region is $x = 12$.

An Analogy: The Jury Trial of an Accused Murderer [MD92] (continued)

5. *Assumption*: The jury is assumed to represent a random sample of citizens who have no prejudice concerning the case.
6. *Experiment and calculation of the test statistic*: The sampling experiment is analogous to the jury selection, the trial, and the jury deliberations. The final vote of the jury is analogous to the calculation of the test statistic.

An Analogy: The Jury Trial of an Accused Murderer [MD92] (continued)

7. *Conclusion*:
   - If the jury's vote is unanimous for guilt, H_0 is rejected and the court concludes that the accused is *guilty*. Although the court does not know the probability α that the conclusion is in error, the system relies on the belief that the value is made very small by requiring a unanimous vote before guilt is concluded.
   - Otherwise, the court reserves judgment about the hypotheses, either by declaring the accused *not guilty*, or by declaring a mistrial and repeating the "test" with a new jury. The court never accepts the null hypothesis by declaring the accused *innocent*, perhaps recognizing both that innocence is the status-quo hypothesis and does not need to be proved and that the probability β of incorrectly concluding innocence may not be as small as α.

Observed Significance Levels

[MD92]

The *observed significance level*, or *p-value*, for a specific statistical test is the probability (assuming H_0 is true) of observing a value of the test statistic that is at least as contradictory to H_0, and supportive of H_a, as the one computed from the sample data.

The closer the p-value is to zero, the more strongly the test statistic disagrees with H_0.

Reporting Test Results as p-Values

How to Decide Whether to Reject H_0

1. Choose the maximum value of α that you are willing to tolerate.
2. If the observed significance level (p-value) of the test is less than the chosen value of α, then reject H_0.

Calculating the *p*-Value for a Test of Hypothesis [MD92]

1. Determine the value of the test statistic z corresponding to the result of the sampling experiment.
2.
   - If the rejection region is of the form $>$, then the p-value is the tail area to the right of, or above, the observed z value.
   - If the rejection region is of the form $<$, then the p-value is the tail area to the left of, or below, the observed z value.
   - If the rejection region is of the form $\neq$ and the observed z value is positive, then the p-value is equal to twice the tail area to the right of, or above, the observed z value.
   - If the rejection region is of the form $\neq$ and the observed z value is negative, then the p-value is equal to twice the tail area to the left of, or below, the observed z value.

Distribution of Students by College

[CFG08]

A sample of 180 students is taken from a population of 1301 college freshmen using stratified random sampling.

College	N	n
Arts and Sciences	131	18
Commerce	432	59
Computer Studies	202	29
Education	83	11
Engineering	103	14
Nursing	350	49
Total	1301	180

Does the distribution by college for the sample differ significantly from the distribution by college for the population?

Distribution of Students by College

An inferential statistical problem

1. The population of interest is the set of 1301 college freshmen.
2. The variable to be investigated is the distribution by college.
3. The sample is the set of 180 students chosen using stratified random sampling.
4. The inference is that the distribution by college for the sample does not differ significantly from the distribution by college for the population.
5. The measure of reliability for the inference is a level of significance $\alpha = 0.05$.

Properties of a Multinomial Experiment

[MD92]

1. The experiment consists of n identical trials.
2. There are k possible outcomes to each trial.
3. The probabilities of the k outcomes, denoted by p_1, p_2, ..., p_k, remain the same from trial to trial, where $p_1 + p_2 + \cdots + p_k = 1$.
4. The trials are independent.
5. The random variables of interest are the counts n_1, n_2, ..., n_k in each of the k cells.

Chi-Square Goodness-of-Fit Test

[MD92, SC88]

H_0:	$p_1 = p_{1,0}$, $p_2 = p_{2,0}$, ..., $p_k = p_{k,0}$ where $p_{1,0}, p_{2,0}, \ldots, p_{k,0}$ represent the hypothesized values of the multinomial probabilities
H_a:	At least one of the multinomial probabilities does not equal its hypothesized value
Test statistic:	$X^2 = \sum_{i=1}^{k} \frac{[n_i - E(n_i)]^2}{E(n_i)}$ where $E(n_i) = np_{i,0}$, the expected number of outcomes of type i assuming that H_0 is true.
Rejection region:	$X^2 > \chi^2_\alpha$ where χ^2_α has $(k-1)$ degrees of freedom

Chi-Square Goodness-of-Fit Test

[MD92, SC88]

(continued)

Assumptions:

1. A multinomial experiment has been conducted. This is generally satisfied by taking a random sample from the population of interest.
2. The sample size n will be large enough so that for every cell, the expected cell count $E(n_i)$ will be equal to 5 or more.

Distribution of Students by College

A test of hypothesis

1. H_0: $p_1 = \frac{131}{1301}$, $p_2 = \frac{432}{1301}$, $p_3 = \frac{202}{1301}$, $p_4 = \frac{83}{1301}$, $p_5 = \frac{103}{1301}$, and $p_6 = \frac{350}{1301}$
2. H_a: $p_1 \neq \frac{131}{1301}$, $p_2 \neq \frac{432}{1301}$, $p_3 \neq \frac{202}{1301}$, $p_4 \neq \frac{83}{1301}$, $p_5 \neq \frac{103}{1301}$, or $p_6 \neq \frac{350}{1301}$
3. *Test statistic*: $X^2 = \sum_{i=1}^{6} \frac{[n_i - E(n_i)]^2}{E(n_i)}$ where $E(n_i) = 180p_{i,0}$
4. *Rejection region*: $X^2 > \chi^2_{0.05}$ where $\chi^2_{0.05} \approx 11.0705$, $df = 5$
5. *Assumptions*: '$E(n_i) > 5$ for all i' is satisfied.
6. *Experiment and calculation of test statistic*:
   $$X^2 = \frac{[18 - 180(131/1301)]^2}{180(131/1301)} + \cdots + \frac{[49 - 180(350/1301)]^2}{180(350/1301)} \approx 0.0820$$

Distribution of Students by College

A test of hypothesis (continued)

7. *Conclusion*: $X^2 \approx 0.0820$ is not in the rejection region (it is not greater than $\chi^2_{0.05} \approx 11.0705$), so we do not reject H_0. The sample's distribution by college does not differ significantly from the population's distribution by college.

Calculating the *p*-value for a test of hypothesis

1. $X^2 \approx 0.0820$
2. $p = P(\chi^2 \geq X^2) \approx P(\chi^2 \geq 0.0820) \approx 0.99990$ for $df = 5$

Reporting test results as *p*-values

1. $\alpha = 0.05$
2. $p \approx 0.99990$ is not less than $\alpha = 0.05$, so we do not reject H_0.

Distribution of Students by College

An alternative stratified random sample

If the distribution by college for the sample is slightly changed as follows, then $X^2 \approx 0.0334$ with $df = 5$ and $p \approx 0.99999$.

College	N	n
Arts and Sciences	131	18
Commerce	432	**60**
Computer Studies	202	**28**
Education	83	11
Engineering	103	14
Nursing	350	49
Total	1301	180

Distribution of Students by College

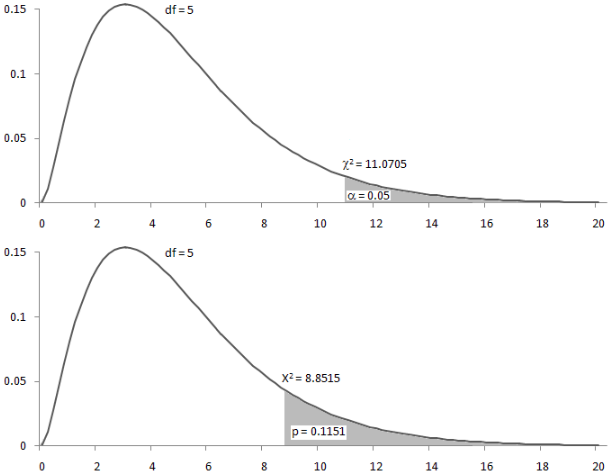
A simple random sample

If simple random sampling was used and we obtained, say, the following, then $X^2 \approx 8.8515$ with $df = 5$ and $p \approx 0.1151$.

College	N	n
Arts and Sciences	131	**9**
Commerce	432	**56**
Computer Studies	202	**27**
Education	83	**10**
Engineering	103	**17**
Nursing	350	**61**
Total	1301	180

Distribution of Students by College

A simple random sample: Graphical illustration



Distribution of Students by College

Microsoft Excel screenshot

	A	B	C	D	E	F
1	College	Population	Predicted	Sample 1	Sample 2	Random
2	Arts and Sciences	131	18.1245196	18	18	9
3	Commerce	432	59.76940815	59	60	56
4	Computer Studies	202	27.94773251	29	28	27
5	Education	83	11.48347425	11	11	10
6	Engineering	103	14.25057648	14	14	17
7	Nursing	350	48.42428901	49	49	61
8	Total	1301		180	180	180
9						
10	df	5				
11	critical value	11.070498	test statistic	0.081985	0.033449	8.851453
12	level of significance	0.05	p-value	0.999901	0.999989	0.115138

Distribution of Students by College

Microsoft Excel screenshot showing formulas

	A	B	C	D	E	F
1	College	Population	Predicted	Sample 1	Sample 2	Random
2	Arts and Sciences	131	=D8*B2/B8	18	18	9
3	Commerce	432	=D8*B3/B8	59	60	56
4	Computer Studies	202	=D8*B4/B8	29	28	27
5	Education	83	=D8*B5/B8	11	11	10
6	Engineering	103	=D8*B6/B8	14	14	17
7	Nursing	350	=D8*B7/B8	49	49	61
8	Total	=SUM(B2:B7)		=SUM(D2:D7)	=SUM(E2:E7)	=SUM(F2:F7)
9						
10	df	=6-1-0				
11	critical value	=CHIINV(B12,B10)	test statistic	=CHIINV(D12,B10)	=CHIINV(E12,B10)	=CHIINV(F12,B10)
12	level of significance	0.05	p-value	=CHITEST(D2:D7,C2:C7)	=CHITEST(E2:E7,C2:C7)	=CHITEST(F2:F7,C2:C7)

Sign Test

[MD92]

Large-sample one-tailed test

H_0:	$M = M_0$
H_a:	$M < M_0$ (or $M > M_0$)
Test statistic:	$z = \frac{(S-0.5)-0.5n}{0.5\sqrt{n}}$ where $S =$ the number of sample measurements less than M_0 (or $S =$ the number of sample measurements greater than M_0 when $H_a : M > M_0$)
Rejection region:	$z > z_\alpha$ where z_α is chosen so that $P(z > z_\alpha) = \alpha$
Assumptions:	The sample is randomly selected from a continuous probability distribution.

Sign Test

[MD92]

Large-sample two-tailed test

H_0:	$M = M_0$
H_a:	$M \neq M_0$
Test statistic:	$z = \frac{(S-0.5)-0.5n}{0.5\sqrt{n}}$ where $S =$ the larger of S_1 and S_2, where S_1 is the number of measurements less than M_0 and S_2 is the number of measurements greater than M_0
Rejection region:	$z > z_{\alpha/2}$ where $z_{\alpha/2}$ is chosen so that $P(z > z_{\alpha/2}) = \alpha/2$
Assumptions:	The sample is randomly selected from a continuous probability distribution.

Binomial Test

[MD92]

Small-sample one-tailed test

H_0:	$M = M_0$
H_a:	$M < M_0$ (or $M > M_0$)
Test statistic:	$S =$ the number of measurements less than M_0 (or $S =$ the number of measurements greater than M_0 when $H_a : M > M_0$)
p-value:	$P(x \geq S)$ where x has a binomial distribution with parameters $n = p = 0.5$
Rejection region:	Reject H_0 if p-value $\leq \alpha$
Assumptions:	The sample is randomly selected from a continuous probability distribution.

Binomial Test

[MD92]

Small-sample two-tailed test

H_0:	$M = M_0$
H_a:	$M \neq M_0$
Test statistic:	$S =$ the larger of S_1 and S_2, where S_1 is the number of measurements less than M_0 and S_2 is the number of measurements greater than M_0
p-value:	$2P(x \geq S)$ where x has a binomial distribution with parameters $n = p = 0.5$
Rejection region:	Reject H_0 if p-value $\leq \alpha$
Assumptions:	The sample is randomly selected from a continuous probability distribution.

Controversy

[Ste51, p. 26, as cited in VW93]

As a matter of fact, most of the scales used widely and effectively by psychologists are ordinal scales. In the strictest propriety the ordinary statistics involving means and standard deviations ought not to be used with these scales . . . On the other hand, . . . there can be invoked a kind of pragmatic sanction: in numerous instances it leads to fruitful results.

z-Test

[MD92]

Large-sample one-tailed test

H_0:	$\mu = \mu_0$
H_a:	$\mu < \mu_0$ (or $\mu > \mu_0$)
Test statistic:	$z = \frac{\overline{x} - \mu_0}{\sigma_{\overline{x}}}$ where $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$
Rejection region:	$z < -z_\alpha$ (or $z > z_\alpha$ when H_a: $\mu > \mu_0$) where z_α is chosen so that $P(z > z_\alpha) = \alpha$
Assumptions:	No assumptions need to be made about the population's probability distribution because the Central Limit Theorem assures us that, for large samples, the test statistic will be approximately normally distributed regardless of the shape of the population's underlying probability distribution.

z-Test

[MD92]

Large-sample two-tailed test

H_0:	$\mu = \mu_0$
H_a:	$\mu \neq \mu_0$
Test statistic:	$z = \frac{\overline{x}-\mu_0}{\sigma_{\overline{x}}}$ where $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$
Rejection region:	$z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$ where $z_{\alpha/2}$ is chosen so that $P(z > z_{\alpha/2}) = \alpha/2$
Assumptions:	No assumptions need to be made about the population's probability distribution because the Central Limit Theorem assures us that, for large samples, the test statistic will be approximately normally distributed regardless of the shape of the population's underlying probability distribution.

t-Test for One Sample

[MD92]

Small-sample one-tailed test

H_0:	$\mu = \mu_0$
H_a:	$\mu < \mu_0$ (or $\mu > \mu_0$)
Test statistic:	$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$
Rejection region:	$t < -t_\alpha$ (or $t > t_\alpha$ when H_a: $\mu > \mu_0$) where t_α is chosen so that $P(t > t_\alpha) = \alpha$ based on $(n-1)$ degrees of freedom
Assumptions:	A random sample is selected from a population with an approximately normal relative frequency distribution.

t-Test for One Sample

[MD92]

Small-sample two-tailed test

H_0:	$\mu = \mu_0$
H_a:	$\mu \neq \mu_0$
Test statistic:	$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$
Rejection region:	$t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$ where $t_{\alpha/2}$ is chosen so that $P(t > t_{\alpha/2}) = \alpha/2$ based on $(n-1)$ degrees of freedom
Assumptions:	A random sample is selected from a population with an approximately normal relative frequency distribution.

Reading Comprehension Performance

[Pad06]

Thirty-one dyslexic and 31 non-dyslexic 11-to-13-year-old Filipino readers of English as a second language are studied.

	Dyslexic		Non-dyslexic	
	$\overline{x}_1$	s_1	$\overline{x}_2$	s_2
Reading comprehension	50.5161	22.0679	53.9667	28.4611
Reading test	13.3226	5.6178	14.7667	6.6731
Analytical reading inventory	37.1936	17.4898	39.2000	22.7663

Does the reading comprehension of dyslexic children differ significantly from that of non-dyslexic children?

t-Test for Two Independent Samples [MD92]

Small-sample two-tailed test

H_0: $\mu_1 - \mu_2 = D_0$

H_a: $\mu_1 - \mu_2 \neq D_0$

Test statistic: $t = \frac{\overline{x}_1 - \overline{x}_2 - D_0}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ where $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$

Rejection region: $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$ where $t_{\alpha/2}$ is chosen so that $P(t > t_{\alpha/2}) = \alpha/2$ based on $(n_1 + n_2 - 2)$ degrees of freedom

Assumptions:

1. Both sampled populations have approximately normal relative frequency distributions.
2. The population variances are equal.
3. The samples are randomly and independently selected from the populations.

Reading Comprehension Performance

A test of hypothesis

1. H_0: $\mu_1 - \mu_2 = 0$
2. H_a: $\mu_1 - \mu_2 \neq 0$
3. *Test statistic*: $t = \frac{\overline{x}_1 - \overline{x}_2 - 0}{\sqrt{s_p^2\left(\frac{1}{31} + \frac{1}{31}\right)}}$ where $s_p^2 = \frac{(31-1)s_1^2 + (31-1)s_2^2}{31+31-2}$
4. *Rejection region*: $t < -t_{0.025}$ or $t > t_{0.025}$ where $t_{0.025} \approx 2.000$, $df = 31 + 31 - 2 = 60$
5. *Assumptions*: The samples are independently selected, but it is not mentioned if the samples are randomly selected.
6. *Experiment and calculation of test statistic*:
   $s_p^2 \approx \frac{(31-1)22.0679^2 + (31-1)28.4611^2}{31+31-2} \approx 648.5132$
   $t \approx \frac{50.5161 - 53.9667 - 0}{\sqrt{648.5132\left(\frac{1}{31} + \frac{1}{31}\right)}} \approx -0.5335$

Reading Comprehension Performance

A test of hypothesis (continued)

7. *Conclusion*: $t \approx -0.5335$ is not in the rejection region (it is not less than $-t_{0.025} \approx -2.000$), so we do not reject H_0. "[T]here is no significant difference (at $p < 0.05$) between the reading comprehension performance of dyslexic children and non-dyslexic children, as measured by both product and process tests." [Pad06, p. 153]

Are the following values [Pad06, p. 153] correct?

	t	p
Reading comprehension	0.53	0.5980
Reading test	0.92	0.3637
Analytical reading inventory	0.39	0.7003

Reading Comprehension Performance

Microsoft Excel screenshot and screenshot showing formulas

	A	B	C	D	E	F
1	df	60		RC	RT	ARI
2	critical value	2.000298	test statistic	0.533459	0.921753	0.389119
3	level of significance	0.05	p-value	0.595687	0.360349	0.698566

	A	B	C	D	E	F
1	df	60		RC	RT	ARI
2	critical value	=TINV(B3,B1)	test statistic	0.533459008	0.921752858	0.389119248
3	level of significance	0.05	p-value	=TDIST(D2,B1,2)	=TDIST(E2,B1,2)	=TDIST(F2,B1,2)

Reading Comprehension Performance

An alternative test of hypothesis

1. H_0: $\mu_1 - \mu_2 = -5$
2. H_a: $\mu_1 - \mu_2 \neq -5$
3. *Test statistic*: $t = \frac{\overline{x}_1 - \overline{x}_2 - (-5)}{\sqrt{s_p^2\left(\frac{1}{31}+\frac{1}{31}\right)}}$ where $s_p^2 = \frac{(31-1)s_1^2+(31-1)s_2^2}{31+31-2}$
4. *Rejection region*: $t < -t_{0.025}$ or $t > t_{0.025}$ where $t_{0.025} \approx 2.000$, $df = 31 + 31 - 2 = 60$
5. *Assumptions*: (the same as before)
6. *Experiment and calculation of test statistic*:
   $s_p^2 \approx \frac{(31-1)22.0679^2+(31-1)28.4611^2}{31+31-2} \approx 648.5132$
   $t \approx \frac{50.5161-53.9667+5}{\sqrt{648.5132\left(\frac{1}{31}+\frac{1}{31}\right)}} \approx 0.2395$

Reading Comprehension Performance

An alternative test of hypothesis (continued)

7. *Conclusion*: $t \approx 0.2395$ is not in the rejection region (it is not greater than $t_{0.025} \approx 2.000$), so we do not reject H_0.

Which H_0 is 'correct,' $\mu_1 - \mu_2 = 0$ or $\mu_1 - \mu_2 = -5$?

They cannot both be correct. This is why we do not (in general) accept H_0. (Accepting is not the same as not rejecting.)

This shows that "hypothesis testing methods can do a good job of proving hypotheses wrong, but they often can't do a very good job of proving hypotheses right." [DC09, p. 236]

Seatwork

The table below [PGRO06, p. 93] shows the results of a test for equality of means between a group with Filipino as the medium of instruction and a group with English as the medium of instruction in a second periodical test in mathematics.

Verify if the values of t, df, and p are correct.

Medium of Instruction	n	$\overline{x}$	s	t	df	p
Filipino	183	13.67	5.9	−0.239	381	0.81
English	200	13.82	5.79			

For $\alpha = 0.05$, is there a significant difference between the Filipino group and the English group in their mathematics achievement?

t-Test for Two Independent Samples [MD92]

Small-sample one-tailed test

H_0:	$\mu_1 - \mu_2 = D_0$
H_a:	$\mu_1 - \mu_2 < D_0$ (or $\mu_1 - \mu_2 > D_0$)
Test statistic:	$t = \frac{\overline{x}_1 - \overline{x}_2 - D_0}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ where $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$
Rejection region:	$t < -t_\alpha$ (or $t > t_\alpha$ when H_a: $\mu_1 - \mu_2 > D_0$) where t_α is chosen so that $P(t > t_\alpha) = \alpha$ based on $(n_1 + n_2 - 2)$ degrees of freedom
Assumptions:	1. Both sampled populations have approximately normal relative frequency distributions. 2. The population variances are equal. 3. The samples are randomly and independently selected from the populations.

Naming Skills Achievement

[Ech07]

The effect of indirect language stimulation on the naming skills of six children with Down syndrome was studied. The table below shows the average naming skills achievement after the treatment.

Child	Low Frequency Stimulation	High Frequency Stimulation
Ann	2.432	2.986
Bea	1.688	3.166
Claire	0.736	1.014
Dana	0.328	0.712
Elan	0.554	0.992
Faye	0.730	0.788

Does the naming skills achievement for low frequency words differ significantly from that for high frequency words?

Naming Skills Achievement

By inspection, it seems that the high frequency treatment yielded significantly better naming skills achievement than the low frequency treatment. However, performing a one-tailed t-test for two independent samples (with $H_0 : \mu_H = \mu_L$ and $H_a : \mu_H > \mu_L$) yields $t \approx 0.9295$, $df = 10$, and $p \approx 0.1873$. If $\alpha = 0.05$, then we do not reject H_0.

But note that the two samples are not independent. Each datum in the first sample is dependent on another datum in the second sample (e.g., 2.432 and 2.986 are for the same child, Ann). The assumption that the samples are independently selected from the populations is not satisfied, and so we cannot use the t-test for two independent samples here.

We use a t-test for paired samples when each datum in the first sample is paired with a datum in the second sample either because the data pair involve the same subject (repeated measures) or because the data pair involve subjects with a matching characteristic (matched pairs).

t-Test for Paired Samples

[MD92]

Small-sample one-tailed test

For each pair of data $x_{1,i}$ and $x_{2,i}$, we create $x_{D,i} = x_{1,i} - x_{2,i}$.

H_0:	$\mu_D = D_0$
H_a:	$\mu_D < D_0$ (or $\mu_D > D_0$)
Test statistic:	$t = \frac{\overline{x}_D - D_0}{s_D/\sqrt{n_D}}$
Rejection region:	$t < -t_\alpha$ (or $t > t_\alpha$ when H_a: $\mu_D > D_0$) where t_α is chosen so that $P(t > t_\alpha) = \alpha$ based on $(n_D - 1)$ degrees of freedom
Assumptions:	1. The population of differences has a normal relative frequency distribution. 2. The differences are randomly selected from the population of differences.

Naming Skills Achievement

A test of hypothesis

1. H_0: $\mu_D = 0$
2. H_a: $\mu_D > 0$
3. *Test statistic*: $t = \frac{\overline{x}_D - 0}{s_D/\sqrt{6}}$
4. *Rejection region*: $t > t_{0.05}$ where $t_{0.05} \approx 2.0150$, $df = 6 - 1 = 5$
5. *Assumptions*: It is not mentioned if the samples are randomly selected.
6. *Experiment and calculation of test statistic*: $t \approx \frac{0.532 - 0}{0.493/\sqrt{6}} \approx 2.64$
7. *Conclusion*: $t \approx 2.64$ is in the rejection region (it is greater than $t_{0.05} \approx 2.0150$), so we reject H_0.
   The naming skills achievement for high frequency words is significantly higher than that for low frequency words.

Naming Skills Achievement

Microsoft Excel screenshot and screenshot showing formulas

	A	B	C
1	Child	Low Freq. Stimulation	High Freq. Stimulation
2	Ann	2.432	2.986
3	Bea	1.688	3.166
4	Claire	0.736	1.014
5	Dana	0.328	0.712
6	Elan	0.554	0.992
7	Faye	0.730	0.788
8	test statistic	0.929544981	2.641512758
9	p-value	0.187250445	0.022946279

	A	B	C
1	Child	Low Freq. Stimulation	High Freq. Stimulation
2	Ann	2.432	2.986
3	Bea	1.688	3.166
4	Claire	0.736	1.014
5	Dana	0.328	0.712
6	Elan	0.554	0.992
7	Faye	0.730	0.788
8	test statistic	=TINV(2*B10,10)	=TINV(2*C10,5)
9	p-value	=TTEST(B2:B7,C2:C7,1,2)	=TTEST(B2:B7,C2:C7,1,1)

t-Test for Paired Samples

[MD92]

Small-sample two-tailed test

For each pair of data $x_{1,i}$ and $x_{2,i}$, we create $x_{D,i} = x_{1,i} - x_{2,i}$.

H_0:	$\mu_D = D_0$
H_a:	$\mu_D \neq D_0$
Test statistic:	$t = \frac{\overline{x}_D - D_0}{s_D/\sqrt{n_D}}$
Rejection region:	$t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$ where $t_{\alpha/2}$ is chosen so that $P(t > t_{\alpha/2}) = \alpha/2$ based on $(n_D - 1)$ degrees of freedom
Assumptions:	1. The population of differences has a normal relative frequency distribution. 2. The differences are randomly selected from the population of differences.

Analysis of Variance

[DC09]

One-way ANOVA for m groups, each with n members

H_0: $\mu_a = \mu_b = \mu_c = \cdots$

H_a: $\mu_a, \mu_b, \mu_c, \cdots$ are not all equal

Test statistic: $F = \frac{ns_*^2}{s^2}$ (see next slide)

Rejection region: $F > F_\alpha$ where F_α is chosen so that $P(F > F_\alpha) = \alpha$ based on $(m-1)$ and $m(n-1)$ degrees of freedom

Assumptions:

1. The groups are independent.
2. The sampled populations have normal relative frequency distributions.
3. The population variances are equal.

Analysis of Variance

[DC09]

Calculating the test statistic F

1. Calculate the sample average for each group:
$\overline{a} = (a_1 + a_2 + \cdots + a_n)/n, \ \ldots$
2. Calculate the average of all the averages:
$\overline{x} = (\overline{a} + \overline{b} + \overline{c} + \cdots)/m$
3. Calculate the sample variance of the averages:
$s_*^2 = [(\overline{a} - \overline{x})^2 + (\overline{b} - \overline{x})^2 + (\overline{c} - \overline{x})^2 + \cdots]/(m-1)$
4. Calculate the sample variance for each group:
$s_a^2 = [(a_1 - \overline{a})^2 + (a_2 - \overline{a})^2 + \cdots + (a_n - \overline{a})^2]/(n-1), \ \ldots$
5. Calculate the average of all the sample variances:
$s^2 = (s_a^2 + s_b^2 + s_c^2 + \cdots)/m$
6. Calculate the value of the F statistic: $F = \frac{ns_*^2}{s^2}$

Analysis of Variance

One-way ANOVA for k groups, with N_1, N_2, ..., N_k members

$$N_1 + N_2 + \cdots + N_k = N$$

$$SS_b = N_1(\overline{x}_1 - \overline{x})^2 + N_2(\overline{x}_2 - \overline{x})^2 + \cdots + N_k(\overline{x}_k - \overline{x})^2$$

$$SS_w = (N_1 - 1)s_1^2 + (N_2 - 1)s_2^2 + \cdots + (N_k - 1)s_k^2$$

Source of variance	Sum of squares	*df*	Mean square	F ratio
Between groups	SS_b	$k-1$	$SS_b/(k-1)$	$\frac{SS_b/(k-1)}{SS_w/(N-k)}$
Within-groups error	SS_w	$N-k$	$SS_w/(N-k)$	
Total	$SS_b + SS_w$	$N-1$		

Seatwork

The tables below [CFG08, p. 41] show the mean grades in ENGS001 of those who passed and those who failed an English Placement Test.

Are the test statistic value and the observed significance level correct? What is H_0 for this test? For $\alpha = 0.05$, should we reject H_0?

EPT	Mean grades in ENGS001	Standard deviation	*N*
Failed	85.33	5.19	167
Passed	88.08	7.01	13
Total	85.53	5.37	180

	Sum of squares	*df*	Mean square	*F*	Sig.
Between groups (combined)	91.052	1	91.052	3.201	0.075
Within groups	5063.809	178	28.448		
Total	5154.861	179			

Simple Linear Regression

[MD92]

Simple linear regression analysis is the methodology of using sample data to estimate and use the straight-line relationship between the mean value of one variable, y, as it relates to a second variable, x.

$$y = \beta_0 + \beta_1 x + \epsilon$$

where

y is the dependent variable (variable to be modeled)
x is the independent variable (variable used as a predictor of y)
ϵ is the random error component
β_0 is the y-intercept of the line
β_1 is the slope of the line

Simple Linear Regression

[MD92]

1. Hypothesize the deterministic component of the model that relates the mean, $E(y)$, to x.
2. Use the sample data to estimate unknown parameters in the model.
3. Specify the probability distribution of ϵ, and estimate the standard deviation of this distribution.
4. Statistically evaluate the usefulness of the model.
5. When satisfied that the model is useful, use it for prediction, estimation, and other purposes.

(We will cover only parts of steps 3 and 4.)

Assessing the Usefulness of the Model [MD92]

Model assumptions

1. The mean of the probability distribution of ϵ is 0.
2. The variance of the probability distribution of ϵ is constant for all settings of x.
3. The probability distribution of ϵ is normal.
4. The values of ϵ associated with any two observed values of y are independent.

Making inferences about the slope

H_0: $\beta_1 = 0$, the linear model contributes no information for the prediction of y

H_a: $\beta_1 \neq 0$, the linear model is useful for predicting y

Inferences about the slope and about the correlation coefficient are similar.

Pearson Product Moment Coefficient of Correlation [MD92]

Small-sample two-tailed test

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{(\sum_{i=1}^{n}(x_i - \overline{x})^2)(\sum_{i=1}^{n}(y_i - \overline{y})^2)}}$$

H_0:	$\rho = 0$
H_a:	$\rho \neq 0$
Test statistic:	$t = \frac{r}{\sqrt{(1-r^2)/(n-2)}}$
Rejection region:	$t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$ where $t_{\alpha/2}$ is chosen so that $P(t > t_{\alpha/2}) = \alpha/2$ based on $(n-2)$ degrees of freedom
Assumptions:	The four assumptions about ϵ

Seatwork

The table below [CFG08, p. 39] shows the correlation of English Placement Test (EPT) results and performance in English language courses of a sample of 180 undergraduates.

Verify if the observed significance levels are correct.

	EPT		
Variable	Pearson correlation	Significance (2-tailed)	Decision
ENGS000	0.460	0.001	Reject H_0
ENGS001	0.353	0.001	Reject H_0
ENGS002	0.121	0.105	Accept H_0

H_0 is that there is no relationship between EPT results and performance in an English language course. For $\alpha = 0.05$, are the decisions above correct?

Ahlyn Combis, Evelyn Florece, and Lydia Goingo.
English Placement Test Results and English Language Performance of Ateneo de Naga University Freshman College Students.
Kamawotan, 2:29–48, 2008.

Cynthia Correo.
Improving the English Proficiency Training Program for AdNU Non-Teaching Personnel Through Assessment.
Kamawotan, 3:3–27, 2009.

Douglas Downing and Jeffrey Clark.
Barron's E-Z Statistics.
Barron's Educational Series, New York, 2009.

Darlene Echavia.
Indirect Language Stimulation and Naming Skills of Children with Down Syndrome.
Education Quarterly, LXV:36–48, 2007.

Josefina Payawal-Gabriel and Marietta Reyes-Otero.
Comprehensible Input Strategies and Pedagogical Moves Using Filipino/English as Medium of Instruction in Secondary Mathematics.
Education Quarterly, LXIV:83–97, 2006.

Qi Guo and Marietta Reyes-Otero.
Language Proficiency and Communication Strategy Preference of Asian Students in the Philippines.
Education Quarterly, LXV:114–128, 2007.

James McClave and Frank Dietrich.
A First Course in Statistics.
Macmillan, New York, 4th edition, 1992.

Joel Noche.
Notes on Descriptive and Inferential Statistics.

Portia Padilla.
Reading Comprehension Performance of Filipino Dyslexic and Non-dyslexic Children.
Education Quarterly, LXIV:148–164, 2006.

Sidney Siegel and N. John Castellan, Jr.
Nonparametric Statistics for the Behavioral Sciences.
McGraw-Hill, New York, 2nd edition, 1988.

Stanley Smith Stevens.
Mathematics, Measurement, and Psychophysics.
In Stanley Smith Stevens, editor, *Handbook of Experimental Psychology*. John Wiley, New York, 1951.

Paul Velleman and Leland Wilkinson.
Nominal, Ordinal, Interval, and Ratio Typologies Are Misleading.
The American Statistician, 47(1):65–72, 1993.

List of Symbols Used

n	sample size (total number of measurements in a sample)
μ_x	population mean of a data set x_i
$\overline{x}$	sample mean of a data set x_i
σ_x	population standard deviation of a data set x_i
s_x	sample standard deviation of a data set x_i
M	median
ρ	population correlation coefficient
r	sample correlation coefficient
$\sum_{i=1}^{n} x_i$	$x_1 + x_2 + \cdots + x_n$