Lecture 10: Essentials of Experimental Design for Interface Evaluation

Part 03

Analysis

Learning Objectives

To provide an overview of the key concepts associated with descriptive and inferential statistics that are often used in interface evaluation.

To provide a description of the key concepts associated with descriptive statistics.

To provide a description of the key concepts associated with inferential statistics.

Learning Outcomes

To be able to define a frequency distribution and variations of the normal curve.

To be able to describe cumulative frequency, percentiles, and measures of central tendency.

To be able to describe standard deviation and standard error.

To be able to describe the z-score and its use.

To be able to describe correlation coefficients and hypothesis testing.

To be able to describe the key aspects of t-tests, Chi-square, and Analysis of Variance, as applied to unrelated and related participant studies.

Sample Article

Below is a sample of the Materials and Methods section taken from a published article: Cassarino, M., Maisto, M., Esposito, Y., Guerrero, D., Chan, J.S. and Setti, A., 2019. Testing attention restoration in a virtual reality driving simulator. *Frontiers in psychology*, *10*, p.250.

Compare understanding pre-and post lecture:

Statistical Analyses:

"Participants' performance at the SART was analyzed in terms of d-prime (d': a measure of signal detection sensitivity, calculated as the standardized difference (z-scores) between the proportion of correct responses on non-lures minus the proportion of incorrect responses on lures), overall mean accuracy (proportion of correct responses on lures and non-lures), mean accuracy on non-lures (pressing the bar), accuracy on lures (not pressing the bar when number three appears), reaction times (in milliseconds) of correct responses (related to pressing the bar in the presence of a non-lure), and inverse efficiency, a measure of speed-accuracy trade-off calculated as the ratio of reaction times over accuracy on non-lures (Bruyer and Brysbaert, 2011). Comparisons between the two exposure groups in terms of gender were conducted using Chi-square test and potential differences in age and driving experience were investigated via an independent samples t-test."

Sample Article

Below is a sample of the Materials and Methods section taken from a published article: Cassarino, M., Maisto, M., Esposito, Y., Guerrero, D., Chan, J.S. and Setti, A., 2019. Testing attention restoration in a virtual reality driving simulator. *Frontiers in psychology*, *10*, p.250.

Compare understanding pre-and post lecture:

Statistical Analyses:

"A 2 × 2 mixed-design ANOVA was conducted with Environment (rural vs. urban) as the between-subjects factor, and SART (pre- vs. post-drive) as the within-subjects factor to investigate effects of environmental exposure on changes in attentional performance pre- and post-drive. Post hoc comparisons were conducted via t-test statistics. Comparisons between exposure groups in terms of driving behavior were assessed via independent t-test. In addition, potential effects of driving on attention were tested through a 2 (SART session) × 2 (environmental exposure) × 2 (driving vs. passenger condition) ANOVA with Driving (driver or passenger) and Environment (urban vs. rural) as the between-subject factors, and SART (pre- vs. post-drive) as the within-subjects factor. We conducted a test of normality on the ANOVA unstandardized residuals as well as the Levene's test of homogeneity; for measures that did not appear to meet the assumptions of normality, we conducted the analyses using non-parametric tests and found no differences in results."

Sample Article

Below is a sample of the Materials and Methods section taken from a published article: Cassarino, M., Maisto, M., Esposito, Y., Guerrero, D., Chan, J.S. and Setti, A., 2019. Testing attention restoration in a virtual reality driving simulator. *Frontiers in psychology*, *10*, p.250.

Compare understanding pre-and post lecture:

Results:

"Environmental Exposure Effects on Attention: The two exposure groups (n = 19 in each group) did not differ significantly in terms of gender (χ21= 0.11, p = 0.74), age (t36 = -0.42, p = 0.67) or driving experience (t36 = 0.16, p = 0.87).

The 2 × 2 mixed-design ANOVA indicated no significant interaction between environmental exposure and SART pre- and post-drive for any of the measures of interest.

There was a main effect of environmental exposure for the measure of d' ($F_{1,36}$ = 4.18, p = 0.048, μ^2 = 0.11), with participants in the rural exposure group (M = 1.26, SD = 1.07) showing overall higher sensitivity (i.e., better performance) than the urban exposure group (M = 0.62, SD = 0.84). There was also a main effect of environmental exposure for the measure of accuracy on lures ($F_{1,36}$ = 4.61, p = 0.04, μ^2 = 0.11), with participants in the rural group (M = 0.64, SD = 0.25) being overall more accurate than those in the urban group (M = 0.48, SD = 0.21). In both cases, however, the size of the effect was small."

Overview

Descriptive statistics: Measures of central tendency, variability.

Inferential statistics: Techniques to allow us to generalise the results from our study sample to the population.

Overview

Descriptive statistics: Measures of central tendency, variability.

Inferential statistics: Techniques to allow us to generalise the results from our study sample to the population.

Descriptive statistics:
Measures of central tendency, variability.

Frequency Distribution

Frequency distribution table, showing experimental scores and the frequency with which the score occurred. N = 18.

Score	Frequency, *f*
17	1
16	0
15	4
14	6
13	4
12	1
11	1
10	1

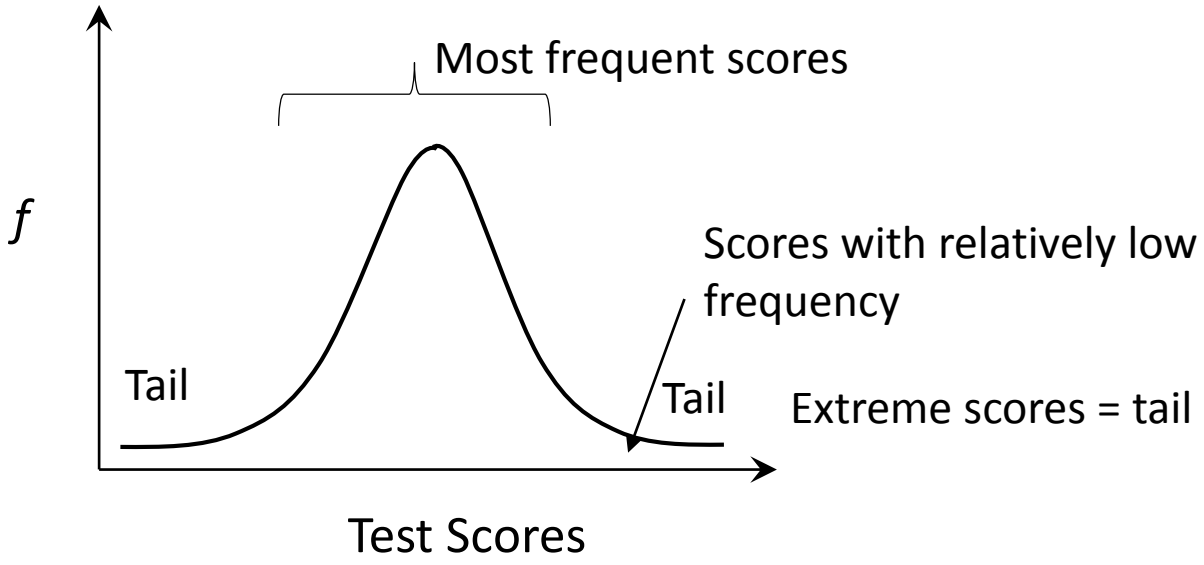
Ages of a sample of people

Can represent as →

Bar graph
Histogram
Frequency polygon

Types of Simple Frequency Distribution

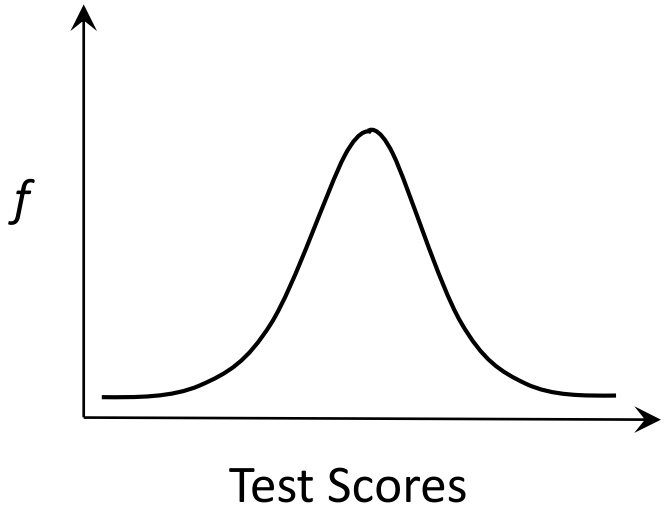
Normal Distribution



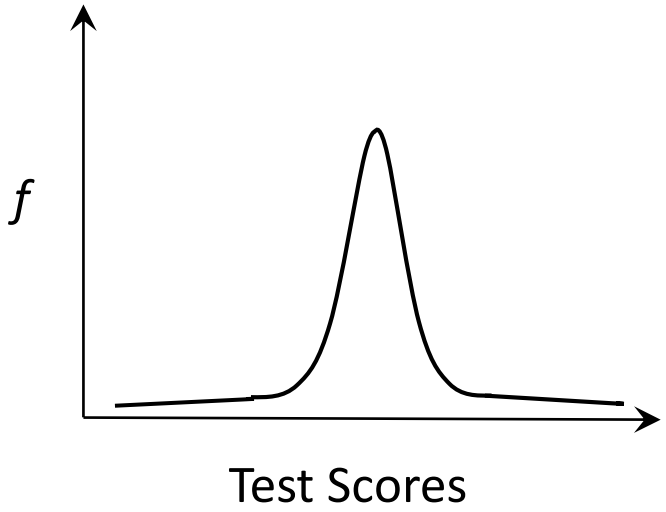
Often apply the normal curve model. i. e., Assume a population of scores comes close enough to forming the normal curve that we treat it as it does form a normal curve and is normally distributed.

Variations in the Normal Curve (*kurtosis*)

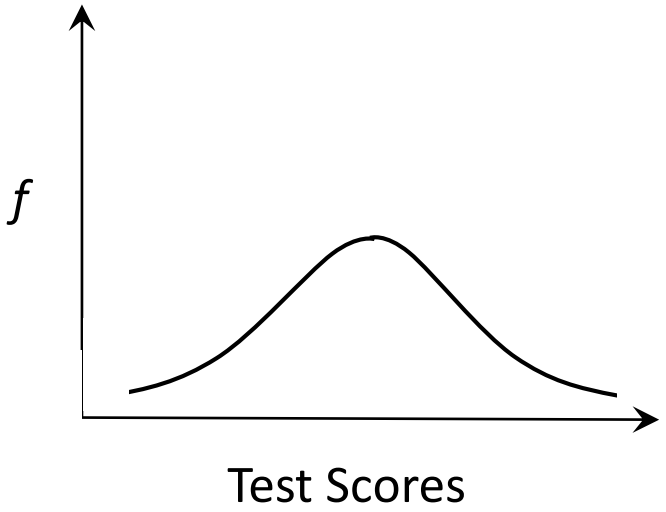
Ideal normal distribution
mesokurtic



Ideal normal distribution
leptokurtic

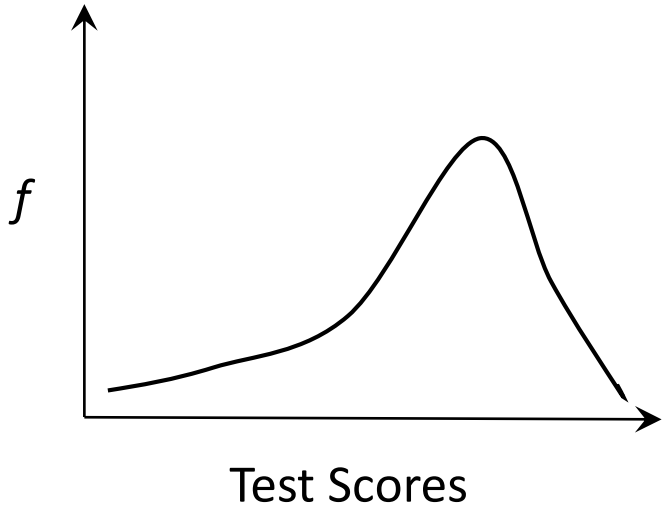


Ideal normal distribution
platykurtic

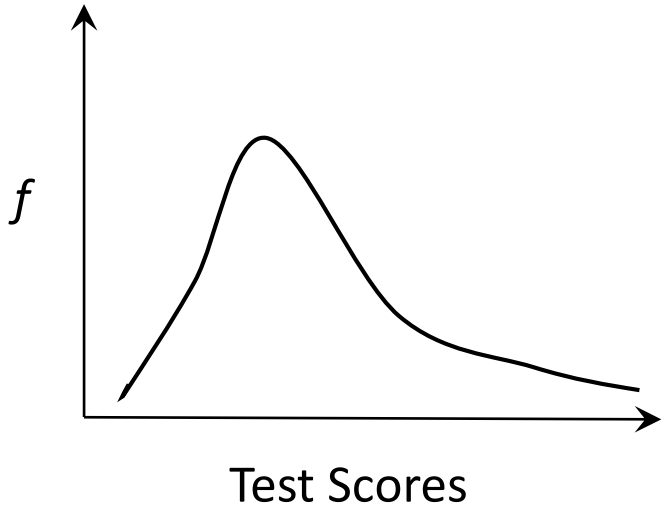


Variations in the Normal Curve (*Skewed distributions*)

Negative skew



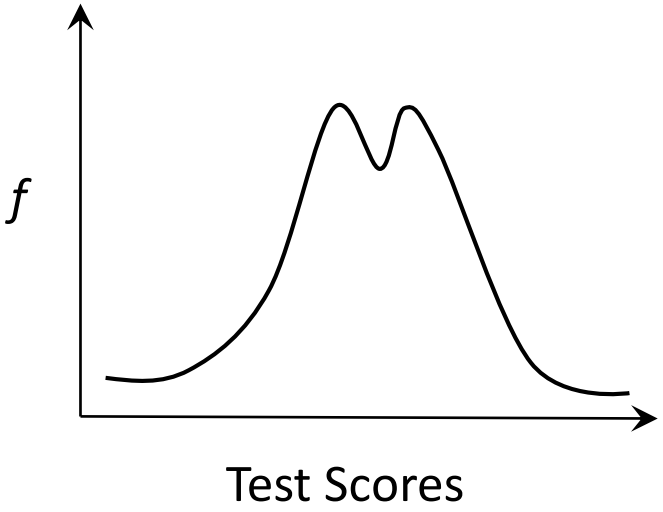
Positive skew



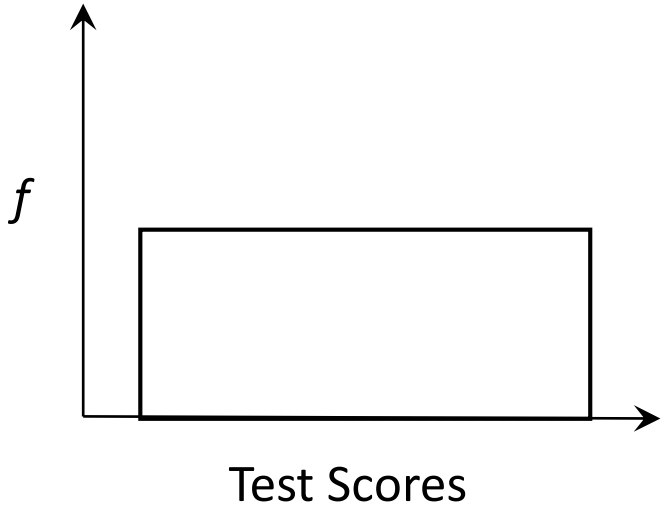
One pronounced tail

Bimodal and Rectangular Distributions

Bimodal



Rectangular



Cumulative Frequency

Frequency and frequency distribution table,

Score	Frequency, *f*	*cf*
17	1	19
16	2	18
15	4	16
14	5	12
13	4	7
12	0	3
11	2	3
10	1	1

Ages of a sample of people

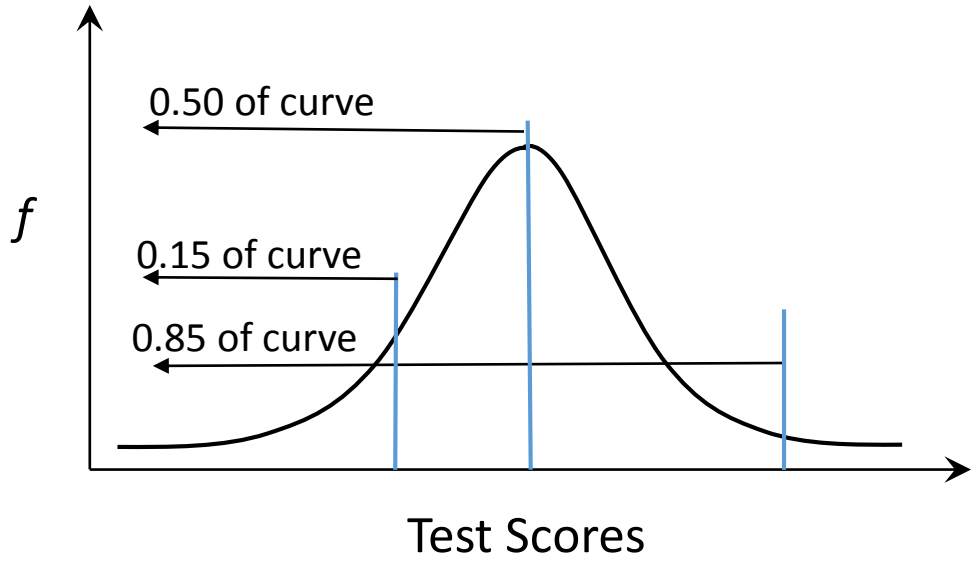
Total = 19

Cumulative frequency- frequency of all scores at or below a particular score.

Relative Frequency = f/N

Percentiles

Normal distribution showing area under the curve to the left of selected scores



Percentile: Percent of all scores in the data that are at or below a certain score.

Measures of Central Tendency

Where is the bulk of the data located?

Mode
Test scores: 2, 3, 3, 4, 4, 4, 4, 5, 5, 6
Mode = 4 (occurs most frequently) Unimodal
In some cases can be bimodal

Median
Test scores: 1, 2, 3, 3, 4, 7, 9, 10, 11
Score at the 50^{th} percentile
Median = score in the middle, 4. (If N= even, median = average of two scores in the middle.

Mean
Test scores: 3, 4, 6, 7. [Arithmetic mean, Sample mean], 20/4 = 5. Can be inaccurate for a skewed distribution.

Variability

Score (ages) *X*	Score – Mean $X - \bar{X}$	Deviation from Mean	Square of deviation from Mean
18	18-20.5	-2.5	6.25
21	21-20.5	0.5	0.25
23	23-20.5	2.5	6.25
18	18-20.5	-2.5	6.25
19	19-20.5	-1.5	2.25
19	19-20.5	-1.5	2.25
19	19-20.5	-1.5	2.25
33	33-20.5	12.5	156.25
18	18-20.5	-2.5	6.25
19	19-20.5	-1.5	2.25
19	19-20.5	-1.5	2.25
20	20-20.5	-0.5	0.25
Total = 246		Total=0	Total=193

Ages (scores; X) from a group.
Mean of scores ($\bar{X}$) = 246/12 = 20.5
Subtract mean from each score - - > deviation from the mean.
Square each deviation from the mean.
Total of squared deviations from the mean/number of scores: 193/12= 16.08 (the variance).

$$\text{Variance} = \frac{\sum(X-\bar{X})^2}{N}$$

Variance estimate:
Best guess as to the variance of a population of scores *if* you only have the data from a small set of scores from that population on which to base your estimate:
Instead of dividing by *N*, divide by *N*-1

Standard Deviation

Standard deviation – "Average" amount that scores differ (deviate) from the mean of the scores.
Standard deviation – Square root of the variance

Score (ages) *X*	Scores squared – Mean $X - \bar{X}$
20	400
25	625
19	361
35	1225
19	361
17	289
15	225
30	900
27	729
$\sum X$ = 207	$\sum X^2$ = 5115

$$\text{Standard Deviation} = \sqrt{\frac{\sum(X-\bar{X})^2}{N}}$$

Estimated standard deviation:
Best guess as to the standard deviation of a population of scores *if* you only have the data from a small set of scores from that population on which to base your estimate:
Instead of dividing by *N*, divide by *N*-1

Standard Error

$$\text{(estimated) standard error} = \frac{\text{(estimated) standard deviation of population}}{\sqrt{N}}$$

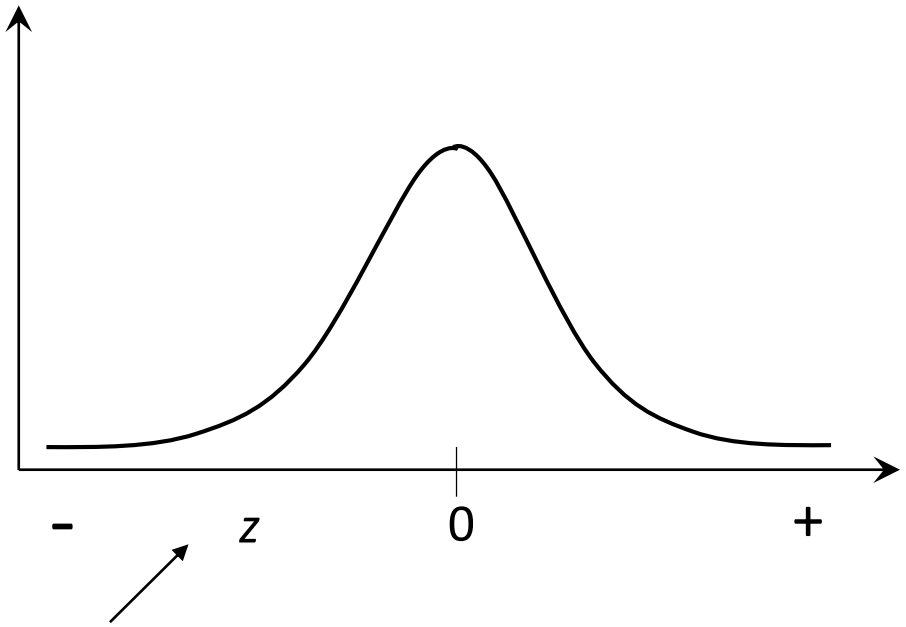
$$\text{(estimated) standard error} = \frac{\sqrt{\frac{\sum X^2 - \frac{(\sum X)2}{N}}{N-1}}}{\sqrt{N}}$$

Standard error: Standard deviation of a number of sample means

Divide standard deviation of scores in the population by the square root of the sample size for which we need to calculate the standard error.

z-scores

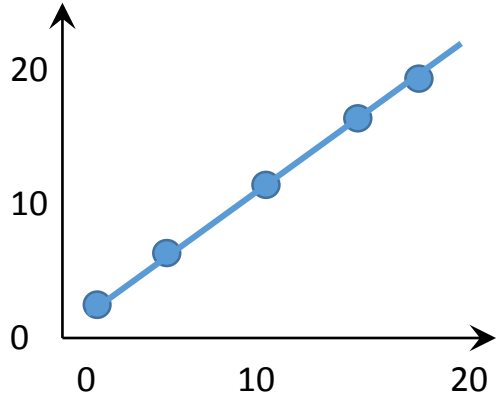
Once size of standard deviation is known, all scores can be re-expressed in terms of the *number of standard deviations they are from the mean*.



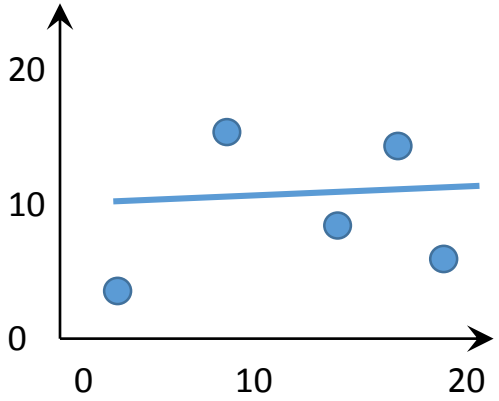
$$z\text{-score} = \frac{X - \bar{X}}{SD}$$

Where,
X is a particular score
$\bar{X}$ is the mean of the set of scores
SD is standard deviation

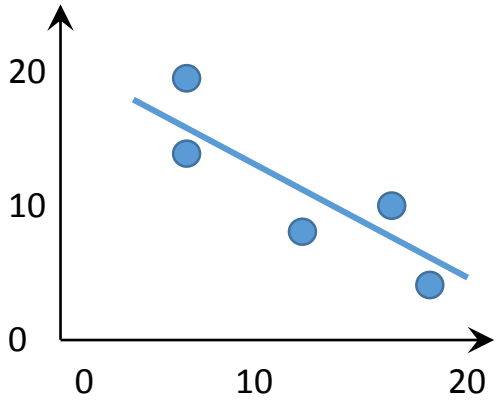
Correlation Coefficients



Near-perfect correlation.
Positive correlation
Close to 1.00



Near-zero correlation.
Close to 0.00

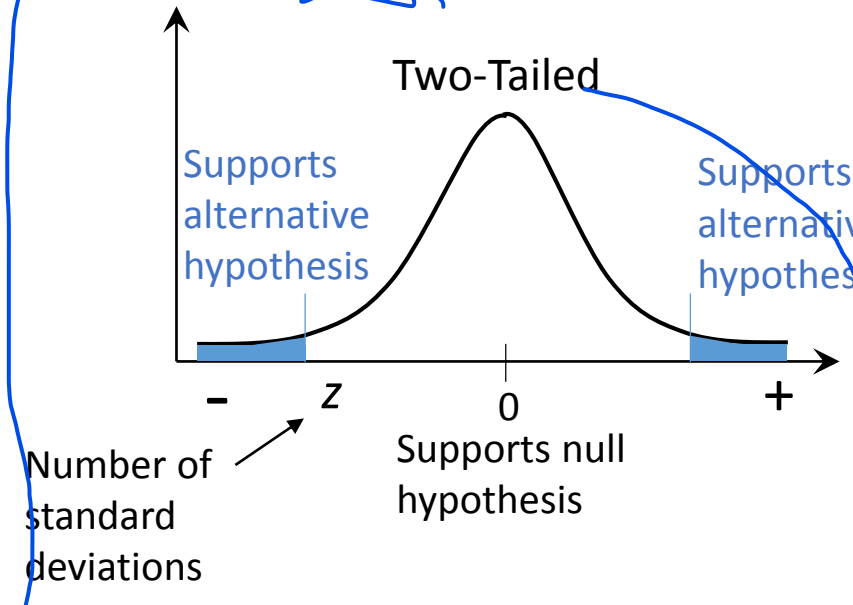


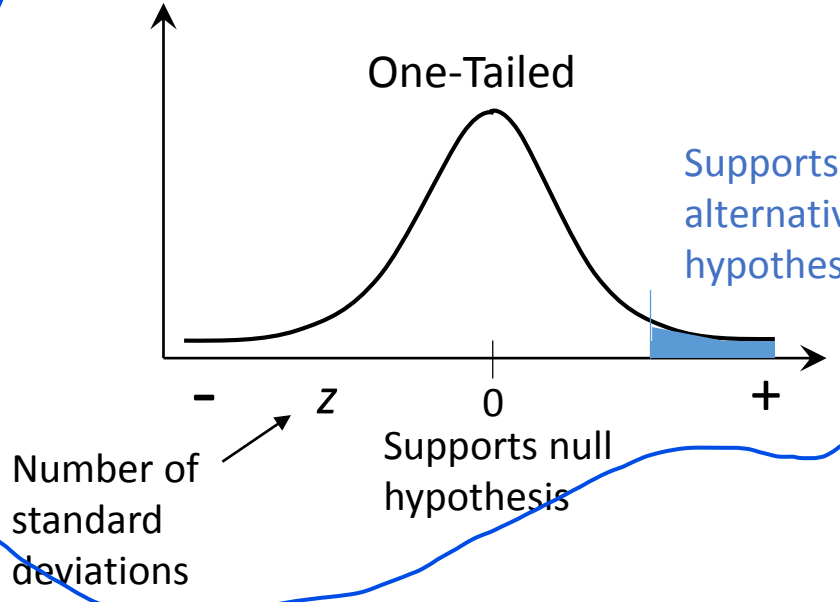
Negative Correlation close to -0.4

Most common correlation coefficient is the Pearson correlation (r):
-Closeness of fit of points to best-fit line
-Slope of fit positive or negative
-A numerical value in range 0.00 to 1.00

r (correlation coefficient) is based on the formula for variance

Hypothesis Testing (significance testing)





If the *null* hypothesis is true then the tested mean should be likely to occur when sampling the raw score population with its own population mean.

If the tested mean occurs rarely (a high z-score), in the tails of the distribution, it is unlikely that the tested mean would occur if sampling the raw score population., and we can *reject the null hypothesis*, and accept the alternative hypothesis.

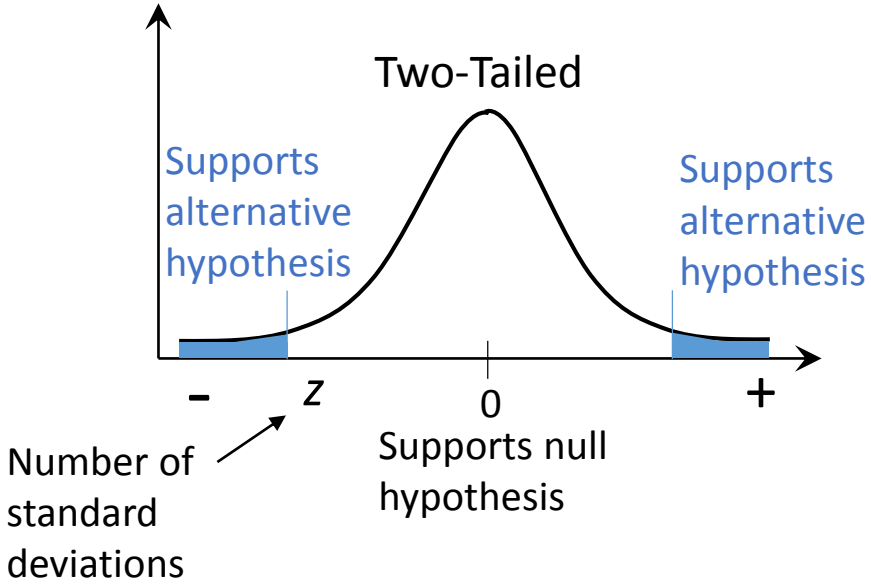
Significant- rejected null hypothesis, and accepted alternative hypothesis.
Not significant- retain null hypothesis.

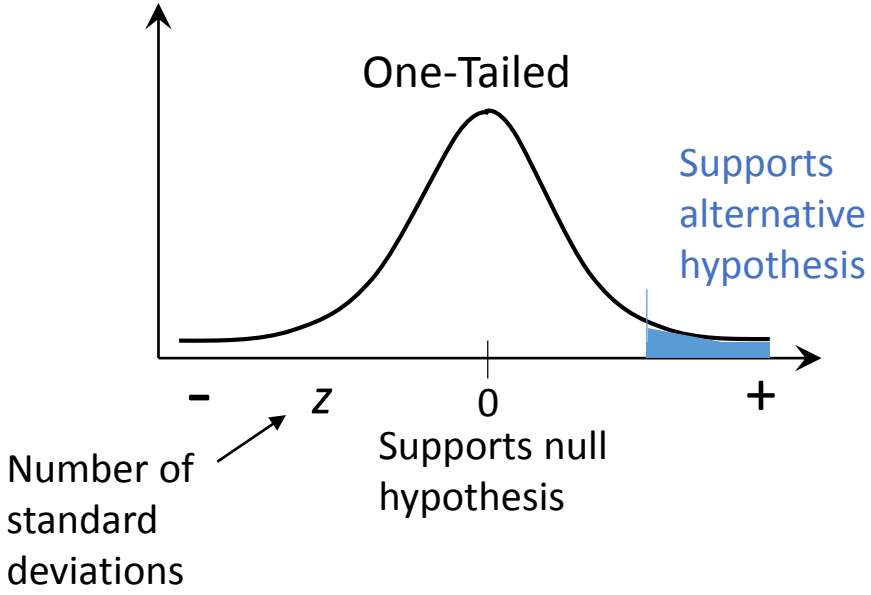
零假设与备择假设 ：
零假设：通常表示没有效应或没有差异的假设 ，例如两组数据没有显著差异。
备择假设：与零假设相对，通常表示研究者期望证明的假设 ，例如存在差异或效应。

单尾检验与双尾检验 ：
双尾检验：当我们想检验两种可能性 （即备择假设可能在零假设的两侧 ）时使用，如图中左右两边的尾部。
单尾检验：只关心一种方向的变化 （正向或负向），如图只考虑一侧的尾部。

显著性水平和 p值：
显著性水平 （通常为0.05）：如果测试结果在 5%的概率内能够发生 ，我们就认为结果是统计显著的 ，可以拒绝零假i
p值：实际观测到的结果 **（或更极端）** 发生的概率。如果 p值小于显著性水平 **（比如小于** 0.05），则拒绝零假设。

Hypothesis Testing (significance testing)





Extreme samples on the distribution.
Check to see if obtained test value lies beyond the critical value.
Region of rejection often taken as 5% (can also use 1%)
If the probability level is less than 0.05, ($p < 0.05$) then the test is statistically significant, can reject null hypothesis (accept alternative hypothesis).
If the probability level is greater than 0.05 ($p > 0.05$), then the test is not statistically significant, and we have to retain the null hypothesis (it is not rejected).

A non-directional hypothesis is two tailed
(e.g., 10 hrs of revision will affect exam performance)
A directional hypothesis is one-tailed
(e.g., 10 hrs of revision will improve exam performance)

Type 1 and Type 2 Errors

Always a chance that there will be an error made deciding to accepting/rejecting a null hypothesis:

A Type 1 error – deciding that the null hypothesis is false (rejecting it) when it is actually true.

A Type 2 error – deciding that the null hypothesis is true when it is actually false

Multiple comparisons: e.g., multiple t-tests.
The more comparisons that are made, more likely is a significant difference due to chance.

Can make an adjustment: Bonferroni correction

Significance level for each test = $\frac{\text{overall significance level}}{\text{number of comparisons}}$