**Harold’s Descriptive Statistics**

**Cheat Sheet**

21 October 2024

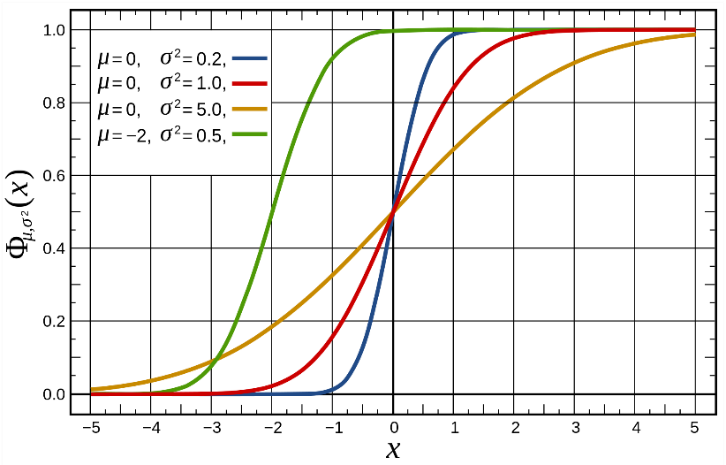
**Descriptive**

|  |  |  |  |
| --- | --- | --- | --- |
| **Description** | **Population** | **Sample** | **Used For** |
| **Data** | Parameters | Statistics | Describing and predicting. |
| **Random Variable** |  |  | The random value from the evaluated population. |
| **Size** |  |  | Number of observations in the population / sample. |

|  |  |  |  |
| --- | --- | --- | --- |
| **Measures of Center** | | (Measure of central tendency) | Indicates which value is typical for the data set. |
| **Mean** |  |  | Measure of center for unordered and frequency distributions. Average, arithmetic mean.  Used when same probabilities for each X. Answers “*Where is the center of the data located?*” |
| **Weighted Mean** |  |  | Some values are counted more than once.  ai = positive integer or percentage. |
| **Median** |  |  | The middle element in a sorted dataset.  More useful when data are skewed with outliers. |
| **Mode** |  | Appropriate for categorical data. | The most frequently-occuring value in a dataset. |
| **Mid-Range** |  | Not often used, easy to compute. | Highly sensitive to unusual values. |
| **Python** | **import** pandas **as** pd  data = pd.read\_csv(‘file.csv’)  print(data.**mean**())  print(data[['Header1']].**median**())  print(data[['Header1', 'Header2']].**mode**())  mid\_range = (data.**min**() + data.**max**()) / 2.0 | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Description** | **Population** | **Sample** | | **Used For** |
| **Measures of Dispersion** | | (Measure of dispersion, variability, or spread of the distribution) | | Reflect the variability of the data (e.g. how different the values are from each other. |
| **Variance** |  |  | | The average of the sum of the square differences.  Not often used. See standard deviation.  Special case of covariance when the two variables are identical. |
| **Covariance** |  |  | | A measure of how much two random variables change together. Measure of “linear dependence”. If X and Y are independent, then their covariance is zero (0). |
| **Standard Deviation** |  |  | | Measure of variation; average distance from the mean. Same units as mean.  Answers “*How spread out is the data?*” |
| **Mean Absolute Deviation** |  |  | | Uses the absolute value instead of the square root of a sum of squares to avoid negative distances. |
| **Pooled Standard Deviation** |  |  | | Inferences for two population means. |
| **Interquartile Range (IQR)** |  |  | | Less sensitive to extreme values. |
| **Range** |  | Not often used, easy to compute. | | Highly sensitive to unusual values. |
| **Python** | **import** pandas **as** pd  data = pd.read\_csv(‘file.csv’)  **print**(data.**var**())  **print**(data.**cov**())  **print**(data.**std**()) | | **print**(data.mad())  def IQR(data): # (import numpy as np)  Q3 = np.**quantile**(data, 0.75)  Q1 = np.**quantile**(data, 0.25)  IQR = Q3 - Q1  range = data.**max**() - data.**min**() | |
| **Description** | **Population** | **Sample** | | **Used For** |
| **Measures of Relative Standing** | | (Measures of relative position) | | Indicates how a particular value compares to the others in the same data set. |
| **Percentile** | Data divided onto 100 equal parts by rank. | | | Important in normal distributions. |
| **Quartile** | Data divided onto 4 equal parts by rank. | | | Used to compute IQR. |
| **Z-Score / Standard Score / Normal Score** |  |  | | The variable measures how many standard deviations the value is away from the mean.  Rule of Thumb: Outlier if . |
| **Calculator (TI-84)** | [2nd][VARS][2] normalcdf(-1**E**99, z) |  | |  |
| **Python** | **import** scipy.stats **as** st  mean, sd, z = 0, 1, 1.5  **print**(st.norm.**cdf**(z, mean, sd)) # P(z **<=** 1.5)  **print**(st.norm.**sf**(z, mean, sd)) # P(z **>=** 1.5)  mean, sd, x = 55, 7.5, 62  **print**(st.norm.**cdf**(x, mean, sd)) # P(x <= 62)  **print**(st.norm.**sf**(x, mean, sd)) # P(x >= 62) | | | 0.9331927987311419  0.0668072012688580  0.8246760551477705  0.1753239448522295 |

Chart, line chart, histogram

Description automatically generated 

**PDF**

**CDF**

|  |  |  |  |
| --- | --- | --- | --- |
| **Example** | **Data** | **Method** | **Results** |
| **Example** | |  |  |
| **Data** | *Unordered Data: 1, 0, 1, 4, 1, 2, 0, 3, 0, 2, 1, 1, 2, 0, 1, 1, 3* | | Relative Frequency: |
| **Manually** | *Ordered Data:*   |  |  | | --- | --- | |  |  | | *0* | *4* | | *1* | *7* | | *2* | *3* | | *3* | *2* | | *4* | *1* | | |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | |  |  |  |  |  |  | | *0* | *4* | *0.35* | *-1.35* | *1.83* | *7.32* | | *1* | *7* | *0.41* | *-0.35* | *0.12* | *0.87* | | *2* | *3* | *0.18* | *0.65* | *0.42* | *1.26* | | *3* | *2* | *0.12* | *1.65* | *2.71* | *5.43* | | *4* | *1* | *0.06* | *2.65* | *7.01* | *7.01* | |  |
| **Calculator (TI-84)** |  | 1. [STAT] [1] selects the list edit screen 2. Move the cursor up to L1 3. [CLEAR] [ENTER] erases L1 4. Repeat for L2 5. Enter data in L1 and data in L2 6. [STAT] 🡪 [1] to select 1-Var Stats 7. [2nd] [1] [ENTER] for L1 8. [2nd] [2] [ENTER] for L2 9. Calculate [ENTER] | **Output:** |
| **Python** | **import** pandas **as** pd  df = pd.DataFrame( [1,0,1,4,1,2,0,3,0,2,1,1,2,0,1,1,3])  **print**(df.**describe**())  **print**()  **print**(df.**std**(ddof=1)) # Sample SD  **print**(df.**std**(ddof=0)) # Population SD  ====================================================  **from** scipy.stats **import** rv\_discrete  x = [0,1,2,3,4,5,6] # Outcomes  p = [0.1,0.2,0.3,0.1,0.1,0.0,0.2] # Prob of outcomes  discrete\_var = **rv\_discrete**(values=(x,p)) # Links x2p  **print**(discrete\_var.**mean**())  **print**(discrete\_var.**std**()) | | **Output:**  count 17.000000  mean 1.352941  std 1.169464  min 0.000000  25% 1.000000  50% 1.000000  75% 2.000000  max 4.000000  std1 1.169464  std0 1.134547 |

**Regression and Correlation**

|  |  |  |
| --- | --- | --- |
| **Description** | **Formula** | **Used For** |
| **Response Variable** |  | Output |
| **Covariate / Predictor Variable** |  | Input |
| **Least-Squares Regression Line** |  | is the slope  is the y-intercept |
| **Regression Coefficient (Slope)** |  | is the slope |
| **Regression Slope Intercept** |  | is the y-intercept |
| **Linear Correlation Coefficient (Sample)** |  | Strength and direction of linear relationship between x and y.  Perfect correlation  Positive linear relationship  Negative linear relationship  No relationship  Strong correlation  Weak correlation  Correlation DOES NOT imply causation. |
| **Residual** |  | Residual = Observed – Predicted |
| **Standard Error of Regression Slope** |  | http://mtweb.mtsu.edu/stats/dictionary/images/defimag/regerrors.gif |
| **Coefficient of Determination** |  | How well the line ﬁts the data.  Represents the percent of the data that is the closest to the line of best fit. Determines how certain we can be in making predictions. |

**Proportions**

|  |  |  |  |
| --- | --- | --- | --- |
| **Description** | **Population** | **Sample** | **Used For** |
| **Proportion** |  |  | Probability of **success**. The proportion of elements that has a particular attribute (x). |
|  |  | Probability of **failure**. The proportion of elements in the population that does not have a specified attribute. |
| **Variance of Population (Sample Proportion)** |  |  | Considered an unbiased estimate of the true population or sample variance. |
| **Pooled Proportion** | *NA* |  | frequency, or number of members in the sample that have the specified attribute. |

**Discrete Random Variables**

|  |  |  |
| --- | --- | --- |
| **Description** | **Formula** | **Used For** |
| **Random Variable** |  | A rule that assigns a number to every **outcome** in the sample space, S.  *e.g.,*  Derived from a probability experiment with different probabilities for each X.  **Used in discrete or finite PDFs.** |
| **Event** |  | An event assigns a value to the random variable X with probability: |
| **Expected Value of *X*** | Each event:  Groups of like events: | E(X) is the same as the mean or average. X takes some countable number of specific values. Discrete.  Expectation of a random variable.  *P(s)* = probability of outcome *s* from *S*. |
| **Linearity of Expectations** |  | When carefully applied, linearity of expectations can greatly simplify calculating expectations.  Does not require that the random variables be independent. |
| **Variance of *X*** |  | Calculate variances with proportions or expected values. |
| **Standard Deviation of *X*** |  | Calculate standard deviations with proportions. |
| **Sum of Probabilities** |  | If same probability, then . |

NOTE: See also “Discrete Definitions” on Harold’s\_Stats\_Distributions\_Cheat\_Sheet.

**Sampling Distribution Statistical Inference**

|  |  |  |
| --- | --- | --- |
| **Description** | **Mean** | **Standard Deviation** |
| **Sampling Distribution** | Is the probability distribution of a statistic; a statistic of a statistic. | |
| **Central Limit Theorem (CLT)** |  | As the sample size drawn from the population with distribution **X** becomes larger, the sampling distribution of the means approaches that of a normal distribution . |
| **Sample Mean** |  | Sampling with replacement:  Sampling without replacement:  (2x accuracy needs 4x n) |
| z-Score |  |  |
| Sample Mean Rule of Thumb | Use if **or** if the population distribution is normal | |
| 10% Condition | . Sample size must be at most 10% of the population size. | |
| **Sample Proportion** |  |  |
| z-Score |  |  |
| Sample Proportion Rule of Thumb | Large Counts Condition:  Use if  Use if | 10 Percent Condition:  Use if |
| **Difference of Sample Means** |  |  |
| Special case when |  |  |
| **Difference of Sample Proportions** |  |  |
| Special case when |  |  |

|  |  |  |
| --- | --- | --- |
| **Bias** | Caused by non-random samples.  Selection Bias: Under coverage, Nonresponse, Voluntary response  Response Bias: Acquiescence, Extreme, Social desirability |  |
| **Variability** | Caused by too small of a sample.  Sampling Methods:  Simple random, systematic, stratified, cluster, convenience |

**Confidence Intervals for One Population Mean / Proportion (σ is Known)**

|  |  |  |
| --- | --- | --- |
| **Description** | **Formula** | |
| **Critical Value (z\*)** | Usually set ahead of time, unless using p-values to determine.  Set to a threshold value of 0.05 (5%) or 0.01 (1%), but always ≤ 0.10 (10%). | |  |  | | --- | --- | | **Confidence Level** | **Critical Value** | | c = 0.90 | z\* = 1.645 | | c = 0.95 | z\* = 1.960 | | c = 0.99 | z\* = 2.576 | |
| **p-value** | Probability of obtaining a sample “more extreme” than the ones observed in your data, assuming *H0* is true. | |
| **Sample Size**  (for estimating µ) | The size of the sample needed to guarantee a confidence interval with a specified margin of error. Rounded up to the nearest whole number. | |
| **Margin of Error / Standard Error (SE)**  (for the estimate of µ) | The estimate differs from the actual value by at most SE.  Use p = 0.50 for worst case if no previous estimate is known. | |
| SE with replacement: | SE without replacement  (with correction factor): |
| **Confidence Interval for µ**  **(z interval)**  ( known, normal population or large sample) |  | |
| **Standardized Test Statistic**  (of the variable from the CLT) |  | |

**Confidence Intervals for One Population Mean / Proportion (σ is Unknown)**

|  |  |  |
| --- | --- | --- |
| **Description** | **Formula** | |
| **Critical Value (t\*)** | Usually set ahead of time, unless using p-values to determine. *df = n-1*.  Set to a threshold value of 0.05 (5%) or 0.01 (1%), but always ≤ 0.10 (10%). | |  |  |  |  | | --- | --- | --- | --- | | ***df*** | **α = 0.10** | **α = 0.05** | **α = 0.01** | | **5** | 2.015 | 2.571 | 4.032 | | **10** | 1.812 | 2.225 | 3.169 | | **15** | 1.753 | 2.131 | 2.947 | | **24** | 1.711 | 2.064 | 2.797 | | **32** | 1.309 | 1.694 | 2.449 | |
| **p-value** | Probability of obtaining a sample “more extreme” than the ones observed in your data, assuming is true. | |
| **Sample Size**  (for estimating µ) | Preliminary estimate of n:  Actual sample size, n:  The size of the sample needed to guarantee a confidence interval with a specified margin of error. Rounded up to the nearest whole number. | |
| **Margin of Error / Standard Error (SE)**  (for the estimate of µ) | The estimate differs from the actual value by at most SE. | |
| SE with replacement: | SE without replacement  (with correction factor): |
| **Confidence Interval for µ**  **(t interval)**  ( unknown, t distribution or small sample) | Chart  Description automatically generated | |
| **Standardized Test Statistic**  (of the variable from the CLT) |  | |

**Confidence Intervals for the Difference Between Two Population Means / Proportions (σ is Known)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Description** | **Formula** | | |
| **Critical Value (z\*)** | Usually set ahead of time, unless using p-values to determine.  Set to a threshold value of 0.05 (5%) or 0.01 (1%), but always ≤ 0.10 (10%). | |  |  | | --- | --- | | **Confidence Level** | **Critical Value** | | c = 0.90 | z\* = 1.645 | | c = 0.95 | z\* = 1.960 | | c = 0.99 | z\* = 2.576 | | |
| **p-value** | TI-84: DISTR 2: normalcdf(z\_test, 99999999) = p | | |
| **Margin of Error / Standard Error (SE)**  (for the estimate of µ) | = Overall probability of success when the two samples are combined.  The estimate differs from the actual value by at most SE.  Use p = 0.50 for worst case if no previous estimate is known. | | |
| **Confidence Interval for µ**  **(z interval)**  ( known, normal population or large sample) |  | | |
| **Standardized Test Statistic**  (of the variable from the CLT) |  | | |
| **Python** | **from** statsmodels.stats.weightstats **import** ztest  sample1 = [21, 28, 40, 55, 58, 60]  sample2 = [13, 29, 50, 55, 71, 90]  **print**(**ztest**(x1 = sample1, x2 = sample2, value = 0)) | | (-0.58017, 0.56179)  z-score = -0.5802  p-value = 0.5618 (two-tailed) |

**Confidence Intervals for the Difference Between Two Population Means / Proportions (σ is Unknown)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Description** | | **Formula** | |
| **Critical Value (t\*)** | | Usually set ahead of time, unless using p-values to determine. *df = n-1*.  Set to a threshold value of 0.05 (5%) or 0.01 (1%), but always ≤ 0.10 (10%). | |  |  |  |  | | --- | --- | --- | --- | | ***df*** | **α = 0.10** | **α = 0.05** | **α = 0.01** | | **5** | 2.015 | 2.571 | 4.032 | | **10** | 1.812 | 2.225 | 3.169 | | **15** | 1.753 | 2.131 | 2.947 | | **24** | 1.711 | 2.064 | 2.797 | | **32** | 1.309 | 1.694 | 2.449 | |
| **p-value** | | TI-84: DISTR 6: tcdf(t\_test, 99999999) = p | |
| **Margin of Error / Standard Error (SE)**  (for the estimate of µ) | | The estimate differs from the actual value by at most SE. | |
| **Confidence Interval for µ**  **(t interval)**  ( unknown, t distribution or small sample) | |  | |
| **Standardized Test Statistic**  (of the variable from the CLT) | | Paired t-test:  Unpaired t-test: | |
| **Python** | **Paired** | **import** scipy.stats **as** st  **import** pandas **as** pd  df = pd.read\_csv(‘ExamScores.csv')  st.ttest\_**rel**(df['Exam1'],df['Exam2']) | Ttest\_relResult  (statistic = 1.4179, pvalue = 0.16254) |
| **Unpaired** | **import** scipy.stats **as** st  **import** pandas **as** pd  df = pd.read\_csv(‘Machine.csv')  st.ttest\_**ind**(df['Old'],df['New'], equal\_var=False)) | Ttest\_indResult  (statistic = 3.3972, pvalue = 0.00324) |
| **from** statsmodels.stats.proportion **import** proportions\_ztest  counts = [95, 125]  n = [5000, 5000]  print(**proportions\_ztest**(counts, n)) | | (statistic = -2.04522, pvalue = 0.04083) |

**Sources:**

* [SNHU MAT-353](https://www.snhu.edu/admission/academic-catalogs/coce-catalog#/courses/HyM37zzl-) - Applied Statistics for STEM, zyBooks.