Harold's Descriptive Statistics
Cheat Sheet

22 October 2022

Descriptive

Description	Population	Sample	Used For
Data	Parameters	Statistics	Describing and predicting.
Random Variable	X, Y	x, y	The random value from the evaluated population.
Size	N	n	Number of observations in the population / sample.

Measures of Center		(Measure of central tendency)	Indicates which value is typical for the data set.
Mean	$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i\, f$ $f = 1\ if\ samples\ are\ unordered$	$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i\, f$ $n = \sum f$	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	$\mu = \frac{\sum a_i\, x_i}{\sum a_i}$	$\bar{\boldsymbol{x}} = \frac{\sum a_i\, x_i}{\sum a_i}$	Some values are counted more than once. a_i = positive integer or percentage.
Median	$Md = \frac{n+1}{2}\ if\ n\ is\ odd$	$Md = \frac{n}{2} + 1\ if\ n\ is\ even$	The middle element in a <u>sorted</u> dataset. More useful when data are skewed with outliers.
Mode	$Mo = \max(f)$	Appropriate for categorical data.	The most frequently-occuring value in a dataset.
Mid-Range	$MidRange = \frac{max. + min.}{2}$	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	$\sigma^2 = \frac{1}{N}\sum (x_i - \mu)^2 f$ $\sigma^2 = \frac{1}{N}\left(\sum_{i=1}^{N} f\, x_i^2 - N\,\mu^2\right)$	$s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2 f$ $s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} f\, x_i^2 - n\,\bar{x}^2\right)$	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	$\sigma(X,Y) = \frac{1}{N}\sum (x - \mu_x)(y - \mu_y)$ $\sigma(X,Y) = \frac{1}{N}\sum_{i=1}^{N} x_i\, y_i - \mu_x\, \mu_y$	$g = \frac{1}{n-1}\sum (x - \bar{x})(y - \bar{y})$ $\sigma(x,y) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i\, y_i - n\,\bar{x}\,\bar{y}\right)$	A measure of how much two random variables change together. Measure of "linear depedence". If X and Y are independent, then their covarience is zero (0).
Standard Deviation	$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$ $\sigma = \sqrt{\frac{\sum x_i^2}{N} - \mu^2}$	$\boldsymbol{s_x} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$ $s = \sqrt{\frac{\sum x_i^2 - n\,\bar{x}^2}{n-1}}$	Measure of variation; average distance from the mean. Same units as mean. Answers "*How spread out is the data?*"
Mean Absolute Deviation	$MAD = \frac{1}{N}\sum \lvert x_i - \mu \rvert$	$MAD = \frac{1}{n}\sum \lvert x_i - \bar{x} \rvert$	Uses the absolute value instead of the square root of a sum of squares to avoid negative distances.
Pooled Standard Deviation	$\sigma_p = \sqrt{\frac{N_1\,\sigma_1^2 + N_2\,\sigma_2^2}{N_1 + N_2}}$	$\boldsymbol{s_p} = \sqrt{\frac{(\boldsymbol{n_1} - \boldsymbol{1})\boldsymbol{s_1^2} + (\boldsymbol{n_2} - \boldsymbol{1})\boldsymbol{s_2^2}}{(\boldsymbol{n_1} - \boldsymbol{1}) + (\boldsymbol{n_2} - \boldsymbol{1})}}$	Inferences for two population means.
Interquartile Range (IQR)	$IQR = Q3 - Q1$	Q1 Q2 Q3 25% | 25% | 25% | 25%	Less sensitive to extreme values.
Range	$Range = max. - min.$	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()
```

<table>
<tr><th>Description</th><th>Population</th><th>Sample</th><th>Used For</th></tr>
<tr><td colspan="2">Measures of Relative Standing</td><td>(Measures of relative position)</td><td>Indicates how a particular value compares to the others in the same data set.</td></tr>
<tr><td>Percentile</td><td colspan="2">Data divided onto 100 equal parts by rank.</td><td>Important in normal distributions.</td></tr>
<tr><td>Quartile</td><td colspan="2">Data divided onto 4 equal parts by rank.</td><td>Used to compute IQR.</td></tr>
<tr><td>Z-Score / Standard Score / Normal Score</td><td>$x = \mu + z\,\sigma$

$z = \frac{x-\mu}{\sigma}$</td><td>$x = \bar{x} + z\,s$

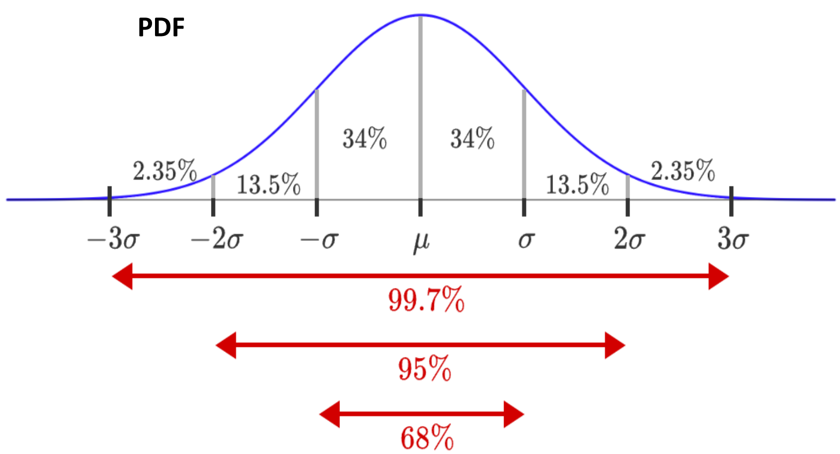
$z = \frac{x-\bar{x}}{s}$</td><td>The z variable measures how many standard deviations the value is away from the mean.
Rule of Thumb: Outlier if $|z| > 2$.</td></tr>
<tr><td>Calculator (TI-84)</td><td>[2nd][VARS][2] normalcdf(-1E99, z)</td><td></td><td></td></tr>
<tr><td>Python</td><td colspan="2"><pre>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)

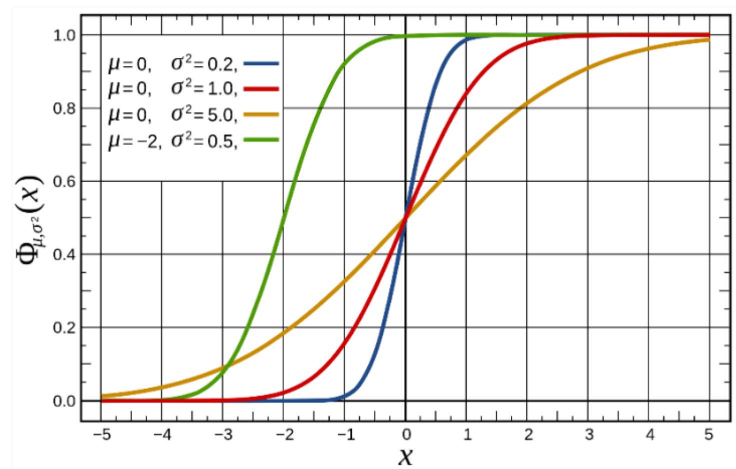
mean, sd, x = 55, 7.5, 62
print(st.norm.cdf(x, mean, sd))   # P(x <= 62)
print(st.norm.sf(x, mean, sd))   # P(x >= 62)</pre></td><td><pre>0.9331927987311419
0.0668072012688580

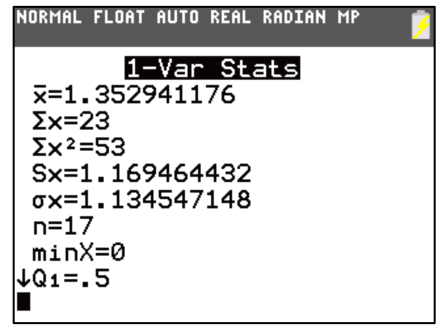
0.8246760551477705
0.1753239448522295</pre></td></tr>
</table>

PDF



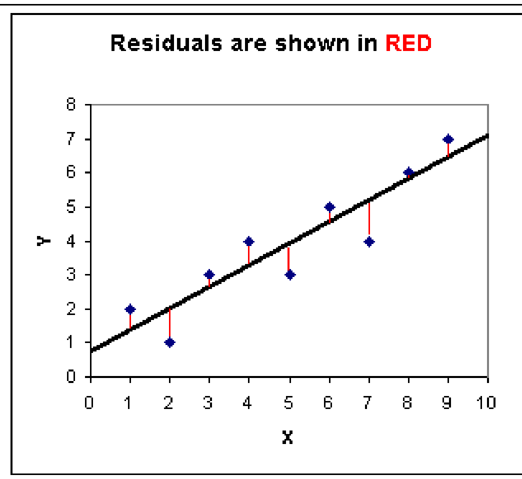
CDF



<table>
<tr><th>Example</th><th>Data</th><th>Method</th><th>Results</th></tr>
<tr><td colspan="2">Example</td><td></td><td></td></tr>
<tr><td>Data</td><td colspan="2">Unordered Data: 1, 0, 1, 4, 1, 2, 0, 3, 0, 2, 1, 1, 2, 0, 1, 1, 3</td><td>$p(x) = f/n$</td></tr>
<tr><td>Manually</td><td>Ordered Data:
<table>
<tr><th>x</th><th>f</th></tr>
<tr><td>0</td><td>4</td></tr>
<tr><td>1</td><td>7</td></tr>
<tr><td>2</td><td>3</td></tr>
<tr><td>3</td><td>2</td></tr>
<tr><td>4</td><td>1</td></tr>
</table></td><td>
<table>
<tr><th>x</th><th>f</th><th>$x - \bar{x}$</th><th>$(x - \bar{x})^2$</th><th>$(x - \bar{x})^2 f$</th></tr>
<tr><td>0</td><td>4</td><td>-1.35</td><td>1.83</td><td>7.32</td></tr>
<tr><td>1</td><td>7</td><td>-0.35</td><td>0.12</td><td>0.87</td></tr>
<tr><td>2</td><td>3</td><td>0.65</td><td>0.42</td><td>1.26</td></tr>
<tr><td>3</td><td>2</td><td>1.65</td><td>2.71</td><td>5.43</td></tr>
<tr><td>4</td><td>1</td><td>2.65</td><td>7.01</td><td>7.01</td></tr>
</table></td><td>$n = \sum f = 4 + 7 + 3 + 2 + 1 = 17$
$\bar{x} = \frac{1}{n}\sum x_i f = \frac{(0 \cdot 4) + \cdots + (4 \cdot 1)}{17} = \frac{23}{17} \approx 1.35$
$\sigma^2 = \frac{1}{n}\sum (x - \bar{x})^2 f = \frac{7.32 + \cdots + 7.01}{17} \approx 1.21$
$\sigma = \sqrt{\sigma^2} \approx 1.13$</td></tr>
<tr><td>Calculator (TI-84)</td><td></td><td>1. [STAT] [1] selects the list edit screen
2. Move cursor up to L1
3. [CLEAR] [ENTER] erases L1
4. Repeat for L2
5. Enter x data in L1 and f 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]</td><td>Output:

</td></tr>
<tr><td>Python</td><td colspan="2"><code>import pandas as pd</code>
<code>df = pd.DataFrame(</code>
<code>[1,0,1,4,1,2,0,3,0,2,1,1,2,0,1,1,3])</code>
<code>print(df.describe())</code>
<code>print()</code>
<code>print(df.std(ddof=1))   # Sample SD</code>
<code>print(df.std(ddof=0))   # Population SD</code>
<code>===</code>
<code>from scipy.stats import rv_discrete</code>
<code>x = [0,1,2,3,4,5,6]   # Outcomes</code>
<code>p = [0.1,0.2,0.3,0.1,0.1,0.0,0.2] # Prob of outcomes</code>
<code>discrete_var = rv_discrete(values=(x,p)) # Links x2p</code>
<code>print(discrete_var.mean())</code>
<code>print(discrete_var.std())</code></td><td>Output:
<code>count   17.000000</code>
<code>mean   1.352941</code>
<code>std   1.169464</code>
<code>min   0.000000</code>
<code>25%   1.000000</code>
<code>50%   1.000000</code>
<code>75%   2.000000</code>
<code>max   4.000000</code>

<code>std1   1.169464</code>
<code>std0   1.134547</code></td></tr>
</table>

Regression and Correlation

Description	Formula	Used For
Response Variable	Y	Output
Covariate / Predictor Variable	X	Input
Least-Squares Regression Line	$\boldsymbol{\hat{y} = b_0 + b_1 x}$	b_1 is the slope b_0 is the y-intercept $(\bar{x}, \bar{y})$ is always a point on the line
Regression Coefficient (Slope)	$\boldsymbol{b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x - \bar{x})^2}}$ $\boldsymbol{b_1 = r\frac{s_y}{s_x}}$	b_1 is the slope
Regression Slope Intercept	$\boldsymbol{b_0 = \bar{y} - b_1\bar{x}}$	b_0 is the y-intercept
Linear Correlation Coefficient (Sample)	$\boldsymbol{r = \frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)}$ $r = \frac{g}{s_x s_y}$	Strength and direction of linear relationship between x and y. $r = \pm 1$ Perfect correlation $r = +0.9$ Positive linear relationship $r = -0.9$ Negative linear relationship $r = \sim 0$ No relationship $r \geq 0.8$ Strong correlation $r \leq 0.5$ Weak correlation Correlation DOES NOT imply causation.
Residual	$\hat{e}_i = y_i - \hat{y}$ $\hat{e}_i = y_i - (b_0 + b_1 x)$ $\sum e_i = \sum(y_i - \hat{y}_i) = 0$	Residual = Observed – Predicted
Standard Error of Regression Slope	$s_{b_1} = \frac{\sqrt{\frac{\sum e_i^2}{n-2}}}{\sqrt{\sum(x_i - \bar{x})^2}}$ $\boldsymbol{s_{b_1} = \frac{\sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}}{\sqrt{\sum(x_i - \bar{x})^2}}}$	
Coefficient of Determination	r^2	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	$P = p = \frac{x}{N}$	$\hat{p} = \frac{x}{n}$	Probability of **success**. The proportion of elements that has a particular attribute (x).
	$q = 1 - p$ $Q = 1 - P$	$\hat{q} = 1 - \hat{p}$	Probability of **failure**. The proportion of elements in the population that does not have a specified attribute.
Variance of Population (Sample Proportion)	$\sigma^2 = \frac{pq}{N}$ $\sigma^2 = \frac{p(1-p)}{N}$	$s_p^2 = \frac{\hat{p}\hat{q}}{n-1}$ $s_p^2 = \frac{\hat{p}(1-\hat{p})}{n-1}$	Considered an unbiased estimate of the true population or sample variance.
Pooled Proportion	*NA*	$\hat{p}_p = \frac{x_1 + x_2}{n_1 + n_2}$ $\hat{p}_p = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$	$x = \hat{p}n =$ frequency, or number of members in the sample that have the specified attribute.

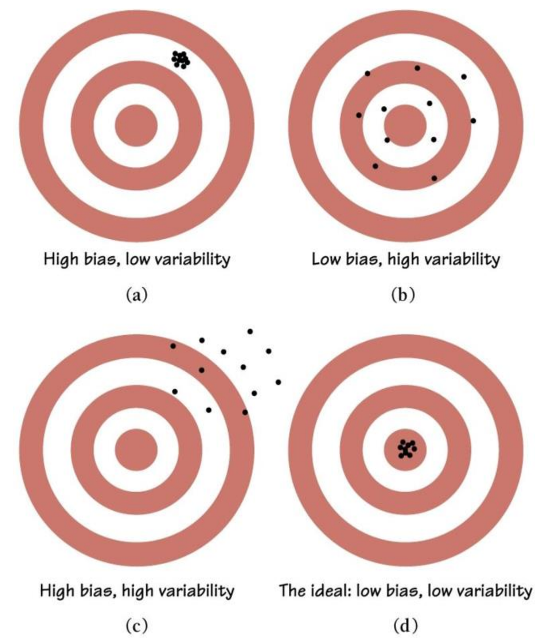
Discrete Random Variables

Description	Formula	Used For
Random Variable	X	A rule that assigns a number to every **outcome** in the sample space, S. $e.g., X(a,b) = a+b=r$ Derived from a probability experiment with different probabilities for each X. **Used in discrete or finite PDFs.**
Event	$X=r$ $X(s)=r$	An event assigns a value to the random variable X with probability: $P(X=r)$
Expected Value of *X*	$E[X]=\bar{x}\ or\ \mu_x$ Each event: $E[X]=\sum P(X)\cdot X$ $E[X]=\sum_{s\in S} X(s)\cdot P(s)$ Groups of like events: $E[X]=\sum_{i=1}^{N} p_i(x)\cdot x_i$ $\boldsymbol{E[X]=\sum_{r\in X(S)} r\cdot P(X=r)}$	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	$E[X+Y]=E[X]+E[Y]$ $E[X+Y+Z]=E[X]+E[Y]+E[Z]$ $E[cX]=cE[X]$	When carefully applied, linearity of expectations can greatly simplify calculating expectations. Does not require that the random variables be independent.
Variance of *X*	$\boldsymbol{V(X)=\sigma_x^2=\sum p_i(x)\cdot(x_i-\mu_x)^2}$ $\sigma_x^2=\sum P(X)\cdot(X-E[X])^2$ $\sigma_x^2=\sum X^2\cdot P(X)-E[X]^2$ $\sigma_x^2=E[X^2]-E[X]^2$	Calculate variances with proportions or expected values.
Standard Deviation of *X*	$SD(X)=\sqrt{V(X)}$ $\sigma_x=\sqrt{\sigma_x^2}$	Calculate standard deviations with proportions.
Sum of Probabilities	$\sum_{i=1}^{N} p_i(x)=1$	If same probability, then $p_i(x)=\frac{1}{N}$.

NOTE: See also "Discrete Definitions" on Harold's Stats Distributions Cheat Sheet.

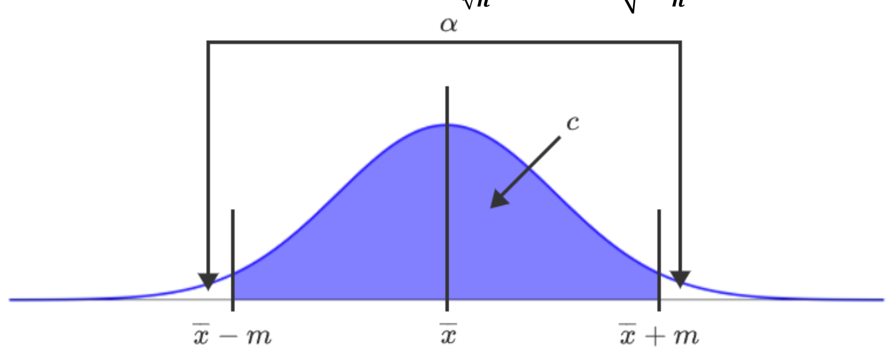
Sampling Distribution Statistical Inference

Description	Mean	Standard Deviation
Sampling Distribution	Is the probability distribution of a statistic; a statistic of a statistic.	
Central Limit Theorem (CLT)	$PDF(\bar{x}) \approx \mathcal{N}\left(0, \frac{\sigma^2}{n}\right)$	As the sample size drawn from the population with distribution **X** becomes larger, the sampling distribution of the means $\overline{\boldsymbol{X}}$ approaches that of a normal distribution $\mathcal{N}\left(0, \frac{\sigma^2}{n}\right)$.
Sample Mean	$\mu_{\bar{x}} = \mu$	Sampling with replacement: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ Sampling without replacement: $\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}} \cdot \frac{\sigma}{\sqrt{n}}$ (2x accuracy needs 4x n)
z-Score	$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$	$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$
Sample Mean Rule of Thumb	Use if $n \geq 30$ **or** if the population distribution is normal	
10% Condition	$n \leq \frac{N}{10}$. Sample size must be at most 10% of the population size.	
Sample Proportion	$\mu = p$	$\sigma_p = \sqrt{\frac{p(1-p)}{n}}$
z-Score	$z = \frac{\hat{p} - \mu}{\sigma_p}$	$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$
Sample Proportion Rule of Thumb	Large Counts Condition: Use if $np \geq 5$ **and** $n(1-p) \geq 5$ Use if $np \geq 10$ **and** $n(1-p) \geq 10$	10 Percent Condition: Use if $N \geq 10n$
Difference of Sample Means	$E(\bar{x}_1 - \bar{x}_2) = \mu_{\bar{x}_1} - \mu_{\bar{x}_2}$	$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
Special case when $\sigma_1 = \sigma_2$		$\sigma_{\bar{x}_1 - \bar{x}_2} = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
Difference of Sample Proportions	$\Delta\hat{p} = \hat{p}_1 - \hat{p}_2$	$\sigma = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
Special case when $p_1 = p_2$		$\sigma = \sqrt{p(1-p)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

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. $n < 30$ Sampling Methods: Simple random, systematic, stratified, cluster, convenience	

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

<table>
<tr><th>Description</th><th colspan="2">Formula</th></tr>
<tr><td>Critical Value (z*)</td><td>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%).</td><td>
<table>
<tr><th>Confidence Level</th><th>Critical Value</th></tr>
<tr><td>c = 0.90</td><td>z* = 1.645</td></tr>
<tr><td>c = 0.95</td><td>z* = 1.960</td></tr>
<tr><td>c = 0.99</td><td>z* = 2.576</td></tr>
</table></td></tr>
<tr><td>p-value</td><td colspan="2">Probability of obtaining a sample "more extreme" than the ones observed in your data, assuming H_0 is true.</td></tr>
<tr><td>Sample Size
(for estimating μ)</td><td colspan="2">$$n = \left(\frac{z^* \sigma}{SE}\right)^2 = \left(\frac{z^*}{SE}\right)^2 p(1-p)$$
The size of the sample needed to guarantee a confidence interval with a specified margin of error. Rounded up to the nearest whole number.</td></tr>
<tr><td rowspan="2">Margin of Error / Standard Error (SE)
(for the estimate of μ)</td><td colspan="2">$$SE(\bar{x}) = m = z^* \frac{\sigma}{\sqrt{n}} = z^* \sqrt{\frac{p(1-p)}{n}}$$
The estimate $\bar{x}$ differs from the actual value by at most SE.
Use p = 0.50 for worst case if no previous estimate is known.</td></tr>
<tr><td>SE with replacement:
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}$$</td><td>SE without replacement
(with correction factor):
$$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}} \cdot \frac{\sigma}{\sqrt{n}}$$</td></tr>
<tr><td>Confidence Interval for μ
(z interval)
(σ known, normal population or large sample)</td><td colspan="2">$$z\ interval = statistic \pm (critical\ value) \bullet (SD\ of\ statistic)$$
$$z\ interval = \bar{x} \pm SE(\bar{x})$$
$$\bar{x} \pm m = [\bar{x} - m,\ \ \bar{x} + m]$$
$$\boldsymbol{z\ interval} = \bar{\boldsymbol{x}} \pm \boldsymbol{z}^* \frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}} = \bar{\boldsymbol{x}} \pm \boldsymbol{z}^* \sqrt{\frac{\boldsymbol{p(1-p)}}{\boldsymbol{n}}}$$


$$\frac{\alpha}{2} = \frac{1-c}{2}$$
$$z^* = z\ score\ for\ probabilities\ of\ {}^{\alpha}/_{2}\ (two-tailed)$$</td></tr>
<tr><td>Standardized Test Statistic
(of the variable $\bar{x}$ from the CLT)</td><td colspan="2">$$z = \frac{statistic - parameter}{SD\ of\ statistic}$$
$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$</td></tr>
</table>

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

<table>
<tr><th>Description</th><th colspan="2">Formula</th></tr>
<tr><td>Critical Value (t*)</td><td>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%).</td><td>

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

</td></tr>
<tr><td>p-value</td><td colspan="2">Probability of obtaining a sample “more extreme” than the ones observed in your data, assuming H_0 is true.</td></tr>
<tr><td>Sample Size
(for estimating μ)</td><td colspan="2">Preliminary estimate of n:

$$n^* = \left(\frac{z^* s}{SE}\right)^2$$

Actual sample size, n:

$$n = \left(\frac{t^* s}{SE}\right)^2$$

The size of the sample needed to guarantee a confidence interval with a specified margin of error. Rounded up to the nearest whole number.</td></tr>
<tr><td rowspan="2">Margin of Error / Standard Error (SE)
(for the estimate of μ)</td><td colspan="2">

$$SE(\bar{x}) = m = t^* \frac{s}{\sqrt{n}}$$

The estimate $\bar{x}$ differs from the actual value by at most SE.</td></tr>
<tr><td>SE with replacement:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

</td><td>SE without replacement
(with correction factor):

$$s_{\bar{x}} = \sqrt{\frac{N-n}{N-1}} \cdot \frac{s}{\sqrt{n}}$$

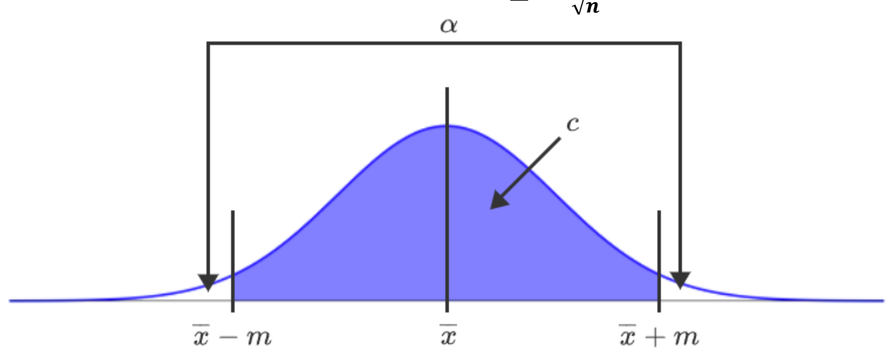
</td></tr>
<tr><td>Confidence Interval for μ
(t interval)
(σ unknown, t distribution or small sample)</td><td colspan="2">

$$t\ interval = statistic\ \pm\ (critical\ value) \bullet (SD\ of\ statistic)$$

$$t\ interval = \bar{x} \pm SE(\bar{x})$$

$$\bar{x} \pm m = [\bar{x} - m,\ \ \bar{x} + m]$$

$$\boldsymbol{t\ interval} = \bar{\boldsymbol{x}}\ \pm\ \boldsymbol{t}^* \frac{\boldsymbol{s}}{\sqrt{\boldsymbol{n}}}$$



</td></tr>
<tr><td>Standardized Test Statistic
(of the variable $\bar{x}$ from the CLT)</td><td colspan="2">

$$t = \frac{statistic - parameter}{SD\ of\ statistic}$$

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

</td></tr>
</table>

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

<table>
<tr><th>Description</th><th>Formula</th></tr>
<tr><td>Critical Value (z*)</td><td>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%).

<table>
<tr><th>Confidence Level</th><th>Critical Value</th></tr>
<tr><td>c = 0.90</td><td>z* = 1.645</td></tr>
<tr><td>c = 0.95</td><td>z* = 1.960</td></tr>
<tr><td>c = 0.99</td><td>z* = 2.576</td></tr>
</table></td></tr>
<tr><td>p-value</td><td>TI-84: DISTR 2: normalcdf(z_test, 99999999) = p</td></tr>
<tr><td>Margin of Error / Standard Error (SE)
(for the estimate of μ)</td><td>

$$E(\bar{x}_1 - \bar{x}_2) = \mu_{\bar{x}1} - \mu_{\bar{x}2}$$

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{SE_1^2 + SE_2^2} = m$$

$$= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = \sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$\hat{p}$ = Overall probability of success when the two samples are combined.
The estimate $\bar{x}_1 - \bar{x}_2$ differs from the actual value by at most SE.
Use p = 0.50 for worst case if no previous estimate is known.</td></tr>
<tr><td>Confidence Interval for μ (z interval)
(σ known, normal population or large sample)</td><td>

$$z\ interval = statistic\ \pm\ (critical\ value) \bullet (SD\ of\ statistic)$$

$$z\ interval = (\bar{x}_1 - \bar{x}_2) \pm SE(\bar{x}_1 - \bar{x}_2)$$

$$(\bar{x}_1 - \bar{x}_2) \pm m = [(\bar{x}_1 - \bar{x}_2) - m,\ \ (\bar{x}_1 - \bar{x}_2) + m]$$

$$\boldsymbol{z\ interval} = (\bar{x}_1 - \bar{x}_2)\ \pm\ \boldsymbol{z^*}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$\boldsymbol{z\ interval} = (\bar{x}_1 - \bar{x}_2)\ \pm\ \boldsymbol{z^*}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

$$\frac{\alpha}{2} = \frac{1-c}{2}$$

$$z^* = z\ score\ for\ probabilities\ of\ {}^{\alpha}/_{2}\ (two - tailed)$$

</td></tr>
<tr><td>Standardized Test Statistic
(of the variable $\bar{x}$ from the CLT)</td><td>

$$z = \frac{observed\ difference - hypothesided\ difference}{SD\ for\ the\ difference}$$

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

</td></tr>
<tr><td>Python</td><td>

```
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))
```

</td><td>(-0.58017, 0.56179)
z-score = -0.5802
p-value = 0.5618
(two-tailed)</td></tr>
</table>

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%).
p-value	TI-84: DISTR 6: tcdf(t_test, 99999999) = p
Margin of Error / Standard Error (SE) (for the estimate of μ)	$SE(\bar{x}_1 - \bar{x}_2) = m = \frac{s_d}{\sqrt{n}}$ The estimate $\bar{x}_1 - \bar{x}_2$ differs from the actual value by at most SE.
Confidence Interval for μ (t interval) (σ unknown, t distribution or small sample)	$t\ interval = statistic\ \pm\ (critical\ value) \bullet (SD\ of\ statistic)$ $\boldsymbol{t\ interval} = (\bar{x}_1 - \bar{x}_2)\ \pm\ \frac{\boldsymbol{s_d}}{\sqrt{\boldsymbol{n}}}$
Standardized Test Statistic (of the variable $\bar{x}$ from the CLT)	$t = \frac{mean\ difference\ between\ samples - parameter}{sample\ SD\ of\ the\ differences\ /\ \sqrt{n}}$ Paired t-test: $t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}}$ $df = n - 1$ Unpaired t-test: $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ $df = n_1 + n_2 - 2$ $t = \frac{\hat{p}_1 - \hat{p}_2}{SE}$

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

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 - Applied Statistics for STEM, zyBooks.