Harold's Statistics
Cheat Sheet

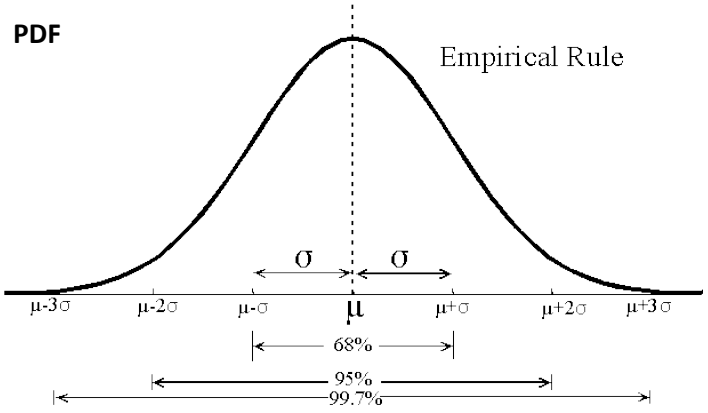
3 Nov 2020

Descriptive

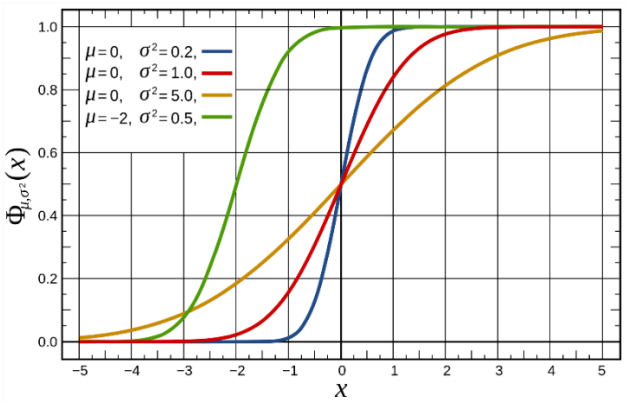
Description	Population	Sample	Used For
Data	Parameters	Statistics	Describing and predicting.
Random Variable	X, Y	x, y	The random value from the evaluated population.
Size	N	n	Number of observations in the population / sample.
Measures Center		(Measure of central tendency)	Indicates which value is typical for the data set.
Mean	$\mu = \frac{1}{N}\sum x_i\, f$ $f = 1\ if\ samples\ are\ unordered$	$\bar{x} = \frac{1}{n}\sum x_i\, f$ $n = \sum f$	Measure of center for unordered and frequency distributions. Average. Includes entire population. Used when same probabilities for each X. Answers "*Where is the center of the data located?*"
Median	$Md = \frac{n+1}{2}\ if\ n\ is\ odd$	More useful when data are skewed.	The middle element in order of rank.
Mode	Mo	Appropriate for categorical data.	The most frequency value in a data set.
Mid-Range	$MidRange = \frac{max. + min.}{2}$	Not often used, easy to compute.	Highly sensitive to unusual values.
Measures Dispersion		(Measure of dispersion or variability)	Reflect the variability of the data (e.g. how different the values are from each other.
Variance	$\sigma^2 = \frac{1}{N}\sum (x-\mu)^2 f$ $\sigma^2 = \frac{1}{N}\left(\sum_{i=1}^{N} f\, x_i^2 - N\,\mu^2\right)$	$s^2 = \frac{1}{n-1}\sum (x-\bar{x})^2 f$ $s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} f\, x_i^2 - n\,\bar{x}^2\right)$	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 depenedence". If X and Y are independent, then their covarience is zero (0).

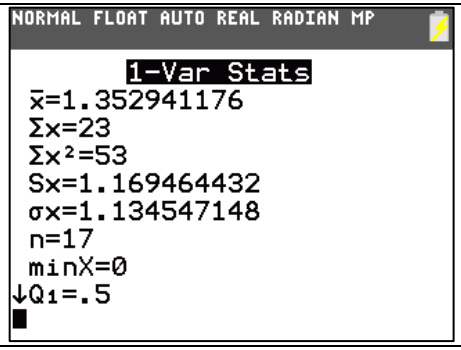
Description	Population	Sample	Used For
Standard Deviation	$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(x-\mu)^2}{N}}$ $\sigma = \sqrt{\frac{\sum x^2}{N} - \mu^2}$	$s_x = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$ $s = \sqrt{\frac{\sum x^2 - n\,\bar{x}^2}{n-1}}$	Measure of variation; average distance from the mean. Same units as mean. Answers "*How spread out is the data?*"
Pooled Standard Deviation	$\sigma_p = \sqrt{\frac{N_1\,\sigma_1^2 + N_2\,\sigma_2^2}{N_1 + N_2}}$	$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)}}$	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.
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	$x = \mu + z\,\sigma$ $z = \frac{x-\mu}{\sigma}$	$x = \bar{x} + z\,s$ $z = \frac{x-\bar{x}}{s}$	The z variable measures how many standard deviations the value is away from the mean. TI-84: [2nd][VARS][2] normalcdf(-1E99, z)

PDF

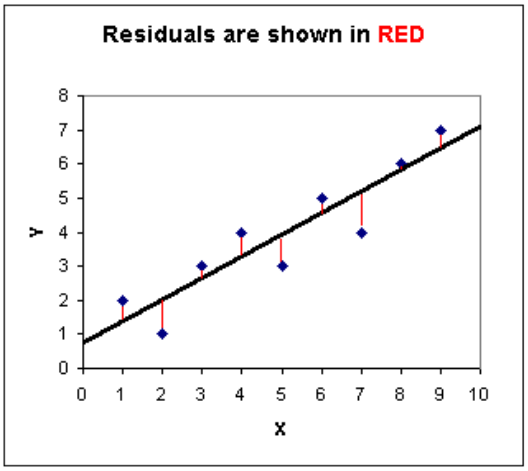


CDF



<table>
<tr><th>Example</th><th>Data</th><th>Method</th><th>Results</th></tr>
<tr><td>Example</td><td colspan="3">Unordered Data: 1, 0, 1, 4, 1, 2, 0, 3, 0, 2, 1, 1, 2, 0, 1, 1, 3</td></tr>
<tr><td>Manually</td><td rowspan="2">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>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 <code>1-Var Stats</code>
7. [2nd] [1] [ENTER] for L1
8. [2nd] [2] [ENTER] for L2
9. Calculate [ENTER]</td><td> 
</td></tr>
</table>

Regression and Correlation

Description	Formula	Used For
Response Variable	Y	Output
Covariate / Predictor Variable	X	Input
Least-Squares Regression Line	$\hat{y} = b_0 + b_1 x$	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)	$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x - \bar{x})^2}$ $b_1 = r\frac{s_y}{s_x}$	b_1 is the slope
Regression Slope Intercept	$b_0 = \bar{y} - b_1 \bar{x}$	b_0 is the y-intercept
Linear Correlation Coefficient (Sample)	$r = \frac{1}{n-1}\sum\left(\frac{x - \bar{x}}{s_x}\right)\left(\frac{y - \bar{y}}{s_y}\right)$ $r = \frac{g}{s_x s_y}$	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	$s_{b_1} = \frac{\sqrt{\frac{\sum e_i^2}{n-2}}}{\sqrt{\sum(x_i - \bar{x})^2}}$ $s_{b_1} = \frac{\sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}}{\sqrt{\sum(x_i - \bar{x})^2}}$	
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.

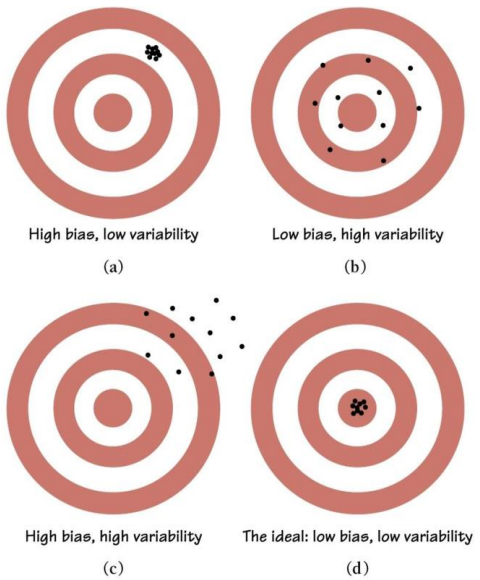
Proportions

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	Derived from a probability experiment with different probabilities for each X. **Used in discrete or finite PDFs.**
Expected Value of *X*	$E(X) = \bar{x}$ $\boldsymbol{E(X) = \mu_x = \sum_{i=1}^{N} p_i\, x_i = \sum P(X)\, X}$	E(X) is the same as the mean. X takes some countable number of specific values. Discrete.
Variance of *X*	$\boldsymbol{Var(X) = \sigma_x^2 = \sum p_i\,(x_i - \mu_x)^2}$ $\sigma_x^2 = \sum P(X)\,(X - E(X))^2$ $\sigma_x^2 = \sum X^2 P(X) - E(X)^2$ $\sigma_x^2 = E(X^2) - E(X)^2$	Calculate variances with proportions or expected values.
Standard Deviation of *X*	$SD(X) = \sqrt{Var(X)}$ $\sigma_x = \sqrt{\sigma_x^2}$	Calculate standard deviations with proportions.
Sum of Probabilities	$\sum_{i=1}^{N} p_i = 1$	If same probability, then $p_i = \frac{1}{N}$.

Statistical Inference

Description	Mean	Standard Deviation
Sampling Distribution	Is the probability distribution of a statistic; a statistic of a statistic.	
Central Limit Theorem (CLT)	$\sqrt{n}(\bar{x}-\mu) \approx \mathcal{N}(0,\sigma^2)$	Lots of $\bar{x}$'s form a Bell Curve, approximating the normal distribution, regardless of the shape of the distribution of the individual x_i's.
Sample Mean	$\boldsymbol{\mu_{\bar{x}} = \mu}$	$\boldsymbol{\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}}$ (2x accuracy needs 4x n)
Sample Mean Rule of Thumb	Use if $n \geq 30$ **or** if the population distribution is normal	
Sample Proportion	$\mu = p$	$\sigma_p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}}$
Sample Proportion Rule of Thumb	Large Counts Condition: Use if $np \geq 10$ **and** $n(1-p) \geq 10$	10 Percent Condition: Use if $N \geq 10n$
Difference of Sample Means	$E(\bar{x}_1 - \bar{x}_2) = \mu_{\bar{x}_1} - \mu_{\bar{x}_2}$	$\boldsymbol{\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
Special case when $\sigma_1 = \sigma_2$		$\boldsymbol{\sigma_{\bar{x}_1-\bar{x}_2} = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
Difference of Sample Proportions	$\Delta\hat{p} = \hat{p}_1 - \hat{p}_2$	$\boldsymbol{\sigma = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$
Special case when $p_1 = p_2$		$\boldsymbol{\sigma = \sqrt{pq}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = \sqrt{p(1-p)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
Bias	Caused by non-random samples.	
Variability	Caused by too small of a sample. $n < 30$	

Confidence Intervals for One Population Mean

Description	Formula
Standardized Test Statistic (of the variable $\bar{x}$)	$$z = \frac{\boldsymbol{statistic - parameter}}{\boldsymbol{standard\ deviation\ of\ statistic}}$$ $$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$
Confidence Interval (C) for μ / z-interval (σ known, normal population or large sample)	$$\boldsymbol{z\text{-}interval = statistic \pm (critical\ value) * (standard\ deviation\ of\ statistic)}$$ $$z\text{ - }interval = \bar{x} \pm E$$ $$= \bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$ $$C + \alpha = 1$$ $$\frac{\alpha}{2} = \frac{1.00 - C}{2}$$ $$z_{\alpha/2} = z - score\ for\ probabilities\ of\ \alpha/2$$
Margin of Error/Standard Error (SE) (for the estimate of μ)	$$E = SE(x) = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$ $$SE(\bar{x}) = s/\sqrt{n}$$
Sample Size (for estimating μ, rounded up)	$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2$$
Critical Value	$z_{\alpha/2}$ Usually set ahead of time, unless using p-values to determine. Usually at a threshold value of 0.05 (5%) or 0.01 (1%), but always ≤ 0.10.
p-value	Probability of obtaining a sample "more extreme" than the ones observed in your data, assuming H_0 is true.