MATH 140
DR. MCLOUGHLIN'S CLASS
STATISTICAL FORMULAE FOR TESTS HANDOUT

$${}_nC_k = \binom{n}{k} = \frac{n!}{k! \cdot (n-k)!}$$

The Binomial variate is the number of successes in n-independent Bernoulli trials where the probability of success at each trial is p. The parameters are p and n (the number of trials).

$p \in (0, 1)$ $n \in \mathbb{N}$ $x \in \{0, 1, 2, \ldots, (n - 1), n\}$

Bin (x, p, n) = $\Pr(X = x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0,1,2,\ldots,n \\ 0 & \text{else} \end{cases}$

$\mu = np$

$\mu_2' = np(np + (1 - p))$

$\mu_3' = np((n-1)(n-2)p^2 + 3p(n - 1) + 1)$

$\sigma^2 = np(1 - p)$

$\mu_3 = np(1 - p)((1 - p) - p)$

$\mu_4 = np((1 + 3p(1-p)(n - 2))$

Let U = S be a well defined universe (the sample space) for our work S will always be able to be a subset of $\mathbb{R}$ (the reals, of course).

Let $D = \{X_1, X_2, X_3, \ldots, X_n\}$ be a finite data set (or we say let $X_1, X_2, X_3, \ldots, X_n$ be a finite random sample for X).

The mode of the sample is a value or values you should know how to find.
The median of the sample is a value or values you should know how to find.

The arithmetic mean of the sample is the value A where $A = \overline{X} = \frac{\sum_{k=1}^{n} X_k}{n}$

The geometric mean of the sample is the value G where $G = \sqrt[n]{\left(\prod_{k=1}^{n} X_k\right)} = \sqrt[n]{X_1 \cdot X_2 \cdot \cdots \cdot X_n}$

for samples such that $X_k > 0$ for all $k \in \mathbb{N}_n$

The harmonic mean of the sample is the value H where $H = \frac{n}{\sum_{i=1}^{n}\left(\frac{1}{X_i}\right)} = \frac{n}{\frac{1}{X_1} + \frac{1}{X_2} + \cdots + \frac{1}{X_n}}$

for samples such that $X_i \neq 0$ for all $i \in \mathbb{N}_n$

The variance, S^2 or S_X^2, is defined as $S^2 = \dfrac{\sum_{k=1}^{n}(X_k - \overline{X})^2}{n-1}$

The standard deviation, S or S_X, is defined as $S = \sqrt{\dfrac{\sum_{k=1}^{n}(X_k - \overline{X})^2}{(n-1)}}$

The range is the (highest value – lowest value)

The inter-quartile range is $Q_3 - Q_1$

The coefficient of variation of the sample is the value C where $C = \dfrac{S}{\overline{X}} \cdot (100\%)$

The mean absolute deviation, MAD, is defined as $MAD = \dfrac{\sum_{k=1}^{n}\left|X_k - \overline{X}\right|}{n}$

Let $D = \{X_1, X_2, X_3, \ldots, X_n\}$ be a finite data set from a population of interest and $C = \{Y_1, Y_2, Y_3, \ldots, Y_n\}$ be a finite data set from a population of interest.
The population parameters for X are the mean, μ_X, the standard deviation, σ_X, etc. ; the population parameters for Y are the mean, μ_Y, the standard deviation, σ_Y, etc.; and, the population (Pearson product-moment) correlation is ρ_{XY}

$$\overline{X} = \frac{\sum_{k=1}^{n} X_k}{n} \qquad \overline{X} \text{ is } \hat{\mu}_X \qquad \overline{Y} = \frac{\sum_{k=1}^{n} Y_k}{n} \qquad \overline{Y} \text{ is } \hat{\mu}_Y$$

$$S_X^2 = \frac{\sum_{k=1}^{n}(X_k - \overline{X})^2}{n-1} \qquad S_X^2 \text{ is } \hat{\sigma}_X^2 \qquad S_Y^2 = \frac{\sum_{k=1}^{n}(Y_k - \overline{Y})^2}{n-1} \qquad S_Y^2 \text{ is } \hat{\sigma}_Y^2$$

$$S_X = \sqrt{\frac{\sum_{k=1}^{n}(X_k - \overline{X})^2}{(n-1)}} \qquad S_X \text{ is } \hat{\sigma}_X \qquad S_Y = \sqrt{\frac{\sum_{k=1}^{n}(Y_k - \overline{Y})^2}{(n-1)}} \qquad S_Y \text{ is } \hat{\sigma}_Y$$

$$r_{xy} = \frac{\sum_{k=1}^{n}(X_k - \overline{X})(Y_k - \overline{Y})}{\sqrt{\sum_{k=1}^{n}(X_k - \overline{X})^2 \sum_{k=1}^{n}(Y_k - \overline{Y})^2}} = \frac{\sum_{k=1}^{n}(X_k - \overline{X})(Y_k - \overline{Y})}{n\sqrt{(S_X)(S_Y)}} = \frac{\sum_{k=1}^{n}(z_{x_k})(z_{y_k})}{n} \qquad r_{xy} = \hat{\rho}_{XY}$$

for X: $Z_i = \dfrac{X_i - \mu_X}{\sigma_X}$ $\quad Z = \dfrac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}}$ $\quad T = \dfrac{X - \overline{X}}{S_X}$ $\quad T_{(n-1)} = \dfrac{\overline{X} - \mu_{\overline{X}}}{S_{\overline{X}}}$

for Y: $Z_i = \dfrac{Y_i - \mu_Y}{\sigma_Y}$ $\quad Z = \dfrac{\overline{Y} - \mu_{\overline{Y}}}{\sigma_{\overline{Y}}}$ $\quad T = \dfrac{Y - \overline{Y}}{S_Y}$ $\quad T_{(n-1)} = \dfrac{\overline{Y} - \mu_{\overline{Y}}}{S_{\overline{Y}}}$

$$\overline{X} = \frac{\sum_{k=1}^{n} X_k}{n} \qquad S_X^2 = \frac{\sum_{k=1}^{n}(X_k - \overline{X})^2}{n-1} \qquad S_X = \sqrt{= \frac{\sum_{k=1}^{n}(X_k - \overline{X})^2}{(n-1)}}$$

$$S_{\overline{X}} = \frac{S_X}{\sqrt{n}} \qquad \sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}} \qquad Z_i = \frac{X_i - \mu_X}{\sigma_X} \qquad Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}}$$

$$T = \frac{X - \overline{X}}{S_X} \qquad T_{n-1} = \frac{\overline{X} - \mu_{\overline{X}}}{S_{\overline{X}}}$$

Independent t-test Pooled Variance Formula

$$t_{pool} = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{\sum_{k=1}^{n}(X_k - \overline{X})^2 + \sum_{k=1}^{n}(Y_k - \overline{Y})^2}{n_1 + n_2 - 2}\right) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$ with df being $n_1 + n_2 - 2$ (‘small’ sample)

$$t_{pool} = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$ with df being $n_1 + n_2 - 2$ (‘large’ sample)

Paired Sample t-test

First compute r_{xy} $$r_{xy} = \frac{\sum_{k=1}^{n}(X_k - \overline{X})(Y_k - \overline{Y})}{\sqrt{\sum_{k=1}^{n}(X_k - \overline{X})^2 \sum_{k=1}^{n}(Y_k - \overline{Y})^2}} = \frac{\sum_{k=1}^{n}(X_k - \overline{X})(Y_k - \overline{Y})}{n\sqrt{(S_X)(S_Y)}} = \frac{\sum_{k=1}^{n}(z_{x_k})(z_{y_k})}{n}$$

Since f the samples are related (two measures from the same subject or matched pairs), the correlated data formula is used.

$$t_{paired} = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{S_X^2}{n}\right) + \left(\frac{S_Y^2}{n}\right) - 2r_{XY} \cdot \left(\frac{S_X \cdot S_Y}{n}\right)}}$$ with df number of pairs minus one.

End, Formulae Handout.