Notation and Equations for Final Exam

Symbol	Definition
X	The variable we measure in a scientific study
n	The size of the sample
N	The size of the population
M	The mean of the sample
μ	The mean of the population (Greek letter mu)
$f(x)$	The frequency of x; the number of scores equal to x
$p(x)$	The probability of x; the fraction of the population for whom $X = x$
σ	Standard deviation of the population (Greek lowercase letter sigma)
σ^2	Variance of the population
$F(x)$	Cumulative distribution
z	z-score
$p(M)$	The distribution of sample means, i.e. the probability distribution for M
σ_M	Standard error of the mean, which equals the standard deviation of $p(M)$
q	The probability of a yes/correct/true outcome for a binary random variable
f or *frequency*	The number of "yes" outcomes in a sample of binary data
H_0	Null hypothesis
H_1	Alternative hypothesis
p	p-value, the probability of getting a result as extreme as you actually got, according to the null hypothesis
α	Alpha level; also the Type I Error Rate
t	t statistic
df	Degrees of freedom
μ_0	Value of the population mean assumed by the null hypothesis in a single-sample t-test
n_A, M_A, μ_A	Sample size, sample mean, and population mean for some group (group A)
$\sigma_{M_A - M_B}$	Standard error of the difference between two sample means (for groups A and B)
t_{crit}	Critical value for t distribution

r	Sample correlation
m	The number of predictor variables in a regression
X_i	A predictor variable in a regression. The subscript i represents any number from 1 through m.
Y	The outcome variable that is being predicted or explained in a regression
$\hat{Y}$ (Y-hat)	The estimated outcome value as predicted by the regression equation
b_i	The regression coefficient for predictor X_i (sometimes written as $b_{\text{predictor name}}$)
b_0	The intercept in the regression equation
σ_{b_i}	The standard error of a regression coefficient
SS_Y	The total sum of squares for the outcome in a regression
$SS_{\text{regression}}$	The sum of squares explained by the predictors in a regression
R^2	The proportion of variability explained by a regression
SS_{total}	The total variability in the data for an ANOVA
$SS_{\text{treatment}}$	Variability explainable by differences among groups (simple ANOVA) or measurements (repeated measures)
SS_{factor}	Variability explainable by the main effect of some factor
$SS_{\text{A:B}}$	Variability explainable by interaction between factors A and B
SS_{residual}	The residual sum of squares, representing the variability that can't be explained in regression or ANOVA
MS_{effect}	Mean square for any effect we might want to test; the subscript can be regression, treatment, Factor, A:B, etc.
df_{effect}	Degrees of freedom for SS_{effect} and MS_{effect}, where *effect* is any effect we might want to test
MS_{residual}	The residual mean square; used as an estimate of the population variance, σ^2 or σ_Y^2
df_{residual}	The degrees of freedom for SS_{residual} and MS_{residual}
F	F statistic
F_{crit}	Critical value for F distribution
k	The number of levels of a factor (treatment) in an ANOVA; written as k_{Factor} when there are multiple factors
$\overline{M}$	The grand mean, i.e. the mean of all the data in all groups taken together
f^{obs}	Observed frequency, in the data
f^{exp}	Expected frequency, based on the null hypothesis
χ^2	Chi-square statistic

Formula name	Formula
Sample mean	$M = \frac{\sum X}{n}$
Mean of finite population	$\mu = \frac{\sum X}{N}$
Expected value or mean of infinite population • R is any random variable, such as raw scores (X) or any statistic (e.g., M) • r represents all values that can occur, and $p(r)$ is the probability of each value	$E(R) = \sum_{r} r \cdot p(r)$
Population variance	$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$
Population standard deviation	$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}}$
Sample variance	$s^2 = \frac{\sum (X - M)^2}{n - 1}$
Sample standard deviation	$s = \sqrt{\frac{\sum (X - M)^2}{n - 1}}$
Cumulative distribution • x represents some value for a raw score • y represents all possible values less than or equal to x	$F(x) = \sum_{y \le x} f(y)$

z-score	$z = \dfrac{X - \mu}{\sigma}$ or $z = \dfrac{X - M}{s}$

- Use the first formula if you know μ and σ and the second if you have to estimate them from the sample using *M* and *s*
- Use the second formula when calculating a correlation

Central Limit Theorem	$p(M) \approx Normal\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

- *p*(*M*) is the probability distribution for *M*, i.e the sampling distribution
- The CLT says this distribution has an approximately normal shape with mean μ and standard deviation $\sigma/\sqrt{n}$
- The normal shape is only guaranteed if the sample size is large enough (rule of thumb: $n \geq 30$)

Conclusions based on p-value	$p > \alpha \rightarrow$ retain H_0 $p < \alpha \rightarrow$ reject H_0, adopt H_1
Standard error of the mean (for single- or paired-samples t-test)	$\sigma_M = \dfrac{s}{\sqrt{n}}$
Effect size for one-sample t-test	$M - \mu_0$
Effect size for independent-samples t-test	$M_A - M_B$
Effect size for paired-samples t-test	M_{diff}
Difference score, for paired-samples t-test	$X_{diff} = X_A - X_B$

t for single sample	$t = \frac{M - \mu_0}{\sigma_M}$
t for independent samples	$t = \frac{M_A - M_B}{\sigma_{M_A - M_B}}$
t for paired samples	$t = \frac{M_{\text{diff}}}{\sigma_{M_{\text{diff}}}}$
p-value for one-tailed t-test predicting positive effect • t_{df} represents the random variable that has a t distribution on *df* degrees of freedom	$p = p(t_{df} \geq t)$
p-value for one-tailed t-test predicting negative effect	$p = p(t_{df} \leq t)$
p-value for two-tailed t-test	$p = 2 \cdot p\left(t_{df} \geq \lvert t \rvert\right)$
Critical value for a two-tailed t-test	$p(t_{df} > t_{crit}) = \alpha/2$
Confidence interval for mean of a single sample	$M \pm t_{crit} \cdot \sigma_M$
Relation between alpha level and confidence level • For example, if you use $\alpha = .01$ to compute t_{crit}, then you end up with a 99% confidence interval	$confidence = 1 - \alpha$
Cohen's *d* for one-sample t-test	$d = \frac{M - \mu_0}{s}$

Correlation	$r = \frac{\sum (z_X \cdot z_Y)}{n-1}$
Interpreting correlation	$r = -1$ → perfect negative relationship $r < 0$ → negative relationship $r = 0$ → no linear relationship $r > 0$ → positive relationship $r = 1$ → perfect positive relationship
Regression equation	$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_m X_m \quad = \quad b_0 + \sum_{i=1 \text{ to } m} b_i X_i$
Total variability in a regression	$SS_Y = \sum (Y - M_Y)^2$
Residual variability in a regression	$SS_{\text{residual}} = \sum (Y - \hat{Y})^2$
Variability explained by a regression	$SS_{\text{regression}} = SS_Y - SS_{\text{residual}}$
Proportion of variability explained by regression	$R^2 = \frac{SS_{\text{regression}}}{SS_Y}$
Explained variability with one predictor	$R^2 = r^2$
t statistic for the i^{th} regression coefficient	$t = \frac{b_i}{\sigma_{b_i}}$

Total sum of squares in an ANOVA	$SS_{\text{total}} = \sum \left(X - \overline{M} \right)^2$
Residual sum of squares in a one-way ANOVA	$SS_{\text{residual}} = \sum \left(X_1 - M_1 \right)^2 + \sum \left(X_2 - M_2 \right)^2 + \ldots + \sum \left(X_k - M_k \right)^2 = \sum_i \left(\sum \left(X_i - M_i \right)^2 \right)$
Treatment sum of squares for one-way ANOVA	$SS_{\text{treatment}} = \sum_i n_i \cdot \left(M_i - \overline{M} \right)^2$
Sum of squares for individual differences in repeated-measures ANOVA	$SS_{\text{subject}} = \sum_s k \cdot \left(M_s - \bar{M} \right)^2$
Mean squares • *source* = regression, treatment, factor, interaction, or residual	$MS_{source} = \frac{SS_{source}}{df_{source}}$
F statistic • *effect* = regression, treatment, factor, or interaction	$F_{effect} = \frac{MS_{effect}}{MS_{\text{residual}}}$
p-value for F test • $F_{df_{effect}, df_{\text{residual}}}$ represents the random variable that has an F distribution on df_{effect} and df_{residual} degrees of freedom	$p = p\left(F_{df_{effect}, df_{\text{residual}}} \geq F_{effect} \right)$
Partitioning variability for regression	$SS_Y = SS_{\text{regression}} + SS_{\text{residual}}$
Partitioning variability for one-way ANOVA	$SS_{\text{total}} = SS_{\text{treatment}} + SS_{\text{residual}}$

Partitioning variability for repeated measures	$SS_{total} = SS_{treatment} + SS_{subject} + SS_{residual}$
Partitioning variability for factorial ANOVA	$SS_{total} = SS_{A} + SS_{B} + SS_{C} + \ldots$ [every main effect] $+ SS_{A:B} + SS_{A:C} + SS_{B:C} + \ldots$ [every 2-way interaction] $+ SS_{A:B:C} + \ldots$ [every possible higher-order interaction] $+ SS_{residual}$
Recognizing an interaction	$M_{a_1,b_1} - M_{a_2,b_1} \neq M_{a_1,b_2} - M_{a_2,b_2}$ → Interaction

- a_1, a_2 are any two levels of Factor A; b_1, b_2 are any two levels of Factor B

Goodness of fit for nominal data	$\chi^2 = \sum \frac{\left(f^{obs} - f^{exp}\right)^2}{f^{exp}}$

- Sum is over all levels of variable (multinomial test) or all combinations of levels of both variables (test of independence)

Expected frequency for multinomial test	$f^{exp}(x) = p(x) \cdot n$

- x is any level of the variable being tested; $p(x)$ is probability of x according to null hypothesis, usually $1/k$ where k is the number of categories

Independence of nominal variables	$p(x \,\&\, y) = p(x) \cdot p(y)$

- x is any level of one variable and y is any level of the other variable

Expected frequency for test of independence	$f^{exp}(x \,\&\, y) = \frac{f^{obs}(x) \cdot f^{obs}(y)}{n}$

- $f^{obs}(x)$ and $f^{obs}(y)$ are the marginal frequencies of x and y

p-value for chi-square tests	$p = p(\chi^2_{df} > \chi^2)$

- χ^2_{df} represents the random variable that has a χ^2 distribution on df degrees of freedom