MA206X — TEE STATISTICS REFERENCE CARD — AY 26-2

Probability, Inference, Regression & ANOVA (Lessons 6–36)

PROBABILITY RULES (Lessons 6–8)

Rule	Formula
Addition	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Complement	$P(A^c) = 1 - P(A)$
Mutually Exclusive	$P(A \cap B) = 0 \implies P(A \cup B) = P(A) + P(B)$
Conditional	$P(A\|B) = \dfrac{P(A \cap B)}{P(B)}$
Multiplication	$P(A \cap B) = P(A\|B) \cdot P(B)$
Independence	$P(A \cap B) = P(A) \cdot P(B)$; also $P(A\|B) = P(A)$
Bayes' Rule	$P(A\|B) = \dfrac{P(B\|A)\,P(A)}{P(B)}$
Total Probability	$P(B) = \sum_i P(B\|A_i)\,P(A_i)$

COUNTING

Multiplication Rule	k stages: $n_1 \cdot n_2 \cdots n_k$
Combinations	$\binom{n}{k} = \dfrac{n!}{k!(n-k)!}$ `choose(n,k)`
Permutations	$P(n,k) = \dfrac{n!}{(n-k)!}$

DISCRETE RANDOM VARIABLES (Lesson 9)

Valid PMF: $p(x) \geq 0$, $\sum_x p(x) = 1$. **CDF:** $F(x) = P(X \leq x) = \sum_{y \leq x} p(y)$

Expected Value	$E(X) = \mu = \sum x \cdot p(x)$
$E[h(X)]$	$\sum h(x) \cdot p(x)$
Variance	$V(X) = \sigma^2 = \sum (x-\mu)^2 p(x)$
Shortcut	$V(X) = E(X^2) - [E(X)]^2$
Std Dev	$\sigma = \sqrt{V(X)}$

NAMED DISCRETE DISTRIBUTIONS (Lessons 10–11)

Binomial	$X \sim \text{Bin}(n, p)$
PMF	$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$, $x = 0, \ldots, n$
$E(X), V(X)$	np, $\quad np(1-p)$
Conditions	Fixed n, two outcomes, constant p, indep. trials
Poisson	$X \sim \text{Pois}(\mu)$
PMF	$p(x) = \dfrac{e^{-\mu}\mu^x}{x!}$, $x = 0, 1, 2, \ldots$
$E(X), V(X)$	μ, $\quad \mu$
Use	Counting events in fixed time/space interval
Scaling	Rate μ scales with interval ($\mu = 5$/hr $\to$ 10/2hr)

CONTINUOUS RANDOM VARIABLES (Lesson 12)

Valid PDF: $f(x) \geq 0$, $\int_{-\infty}^{\infty} f(x)\,dx = 1$. $\quad P(X = c) = 0$

CDF	$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$
Probability	$P(a \leq X \leq b) = \int_a^b f(x)\,dx = F(b) - F(a)$
PDF from CDF	$f(x) = F'(x)$
$E(X)$	$\int_{-\infty}^{\infty} x\,f(x)\,dx$
$E[h(X)]$	$\int_{-\infty}^{\infty} h(x)\,f(x)\,dx$
$V(X)$	$E(X^2) - [E(X)]^2$
Median m	Solve $F(m) = 0.5$

NAMED CONTINUOUS DISTRIBUTIONS (Lessons 13–14)

Normal	$X \sim N(\mu, \sigma^2)$
PDF	$f(x) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$
$E(X), V(X)$	μ, $\quad \sigma^2$
Standardize	$Z = \dfrac{X-\mu}{\sigma} \sim N(0,1)$
$\Phi(z)$	$P(Z \leq z)$ — use `pnorm(z)` in R-lite
Symmetry	$\Phi(-z) = 1 - \Phi(z)$
Percentile	$x = \mu + z\sigma$ where z from `qnorm(p)`
Exponential	$T \sim \text{Exp}(\lambda)$
PDF	$f(t) = \lambda e^{-\lambda t}$, $\quad t \geq 0$
CDF	$F(t) = 1 - e^{-\lambda t}$, $\quad t \geq 0$
$E(T), V(T)$	$\dfrac{1}{\lambda}$, $\quad \dfrac{1}{\lambda^2}$
$P(T > t)$	$e^{-\lambda t}$ (survival function)
Memoryless	$P(T > s+t \mid T > s) = P(T > t)$
Median	$m = \dfrac{\ln 2}{\lambda}$
Connection	Time between events in a Poisson process
Uniform	$X \sim \text{Unif}(a, b)$
PDF	$f(x) = \dfrac{1}{b-a}$, $\quad a \leq x \leq b$
CDF	$F(x) = \dfrac{x-a}{b-a}$
$E(X), V(X)$	$\dfrac{a+b}{2}$, $\quad \dfrac{(b-a)^2}{12}$

R-LITE FUNCTIONS FOR DISTRIBUTIONS

Prefixes: `d`=PMF/PDF, `p`=CDF $P(X \leq x)$, `q`=quantile (inverse CDF)

Dist	Examples
Binomial	`dbinom(x,n,p)`, `pbinom(x,n,p)`, `qbinom(p,n,p)`
Poisson	`dpois(x,lambda)`, `ppois(x,lambda)`
Normal	`pnorm(z)`, `qnorm(p)`; `pnorm(x,mu,sigma)`
Exponential	`pexp(x,rate)`, `qexp(p,rate)`
t	`pt(t,df)`, `qt(p,df)`
F	`pf(F,df1,df2)`, `qf(p,df1,df2)`
Counting	`choose(n,k)`, `factorial(n)`

Right tail: `1 - pdist(...)`.
Two tail: `2*(1 - pdist(abs(stat),...))`.

IDENTIFYING THE DISTRIBUTION

Count successes in n trials $\to$ Binomial
Count events in interval $\to$ Poisson
Time *between* events $\to$ Exponential
Symmetric/bell-shaped measurement $\to$ Normal
Equally likely on $[a, b]$ $\to$ Uniform

STATISTICAL INFERENCE: SAMPLING, HT, AND CIs

SAMPLING DISTRIBUTIONS & CLT (Lesson 17)

Sample mean	$\bar{X} \sim N\left(\mu,\ \dfrac{\sigma^2}{n}\right)$ for large n (CLT)
Standard error	$SE(\bar{X}) = \dfrac{\sigma}{\sqrt{n}}$ (est. $\dfrac{s}{\sqrt{n}}$)
Sample prop.	$\hat{p} \approx N\left(p,\ \dfrac{p(1-p)}{n}\right)$ for large n
CLT conditions	Means: $n \geq 30$ rule of thumb Proportions: $np \geq 10$ and $n(1-p) \geq 10$

HYPOTHESIS TESTING FRAMEWORK (Lesson 20)

H_0	Null hypothesis (status quo, $=$)
H_a	Alternative ($\neq$, $>$, or $<$)
α	Significance level (P of Type I error)
Test statistic	Standardized distance from H_0
p-value	$P(\text{test stat as or more extreme} \mid H_0)$
Decision rule	Reject H_0 if $p < \alpha$; otherwise FTR
Type I (α)	Reject H_0 when true ("false alarm")
Type II (β)	FTR H_0 when false ("missed detection")
Power	$1-\beta$; ↑ with n, α, effect size

CI–HT duality: A two-sided $(1-\alpha)$ CI for a parameter excludes θ_0 *iff* a two-sided level-α test rejects $H_0 : \theta = \theta_0$.

ONE-SAMPLE INFERENCE (Lessons 18–22)

Mean (σ unknown) — One-sample t	
Conditions	Random sample; normal pop or $n \geq 30$ (CLT)
Test stat	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$
CI	$\bar{x} \pm t_{\alpha/2,\, n-1} \cdot \dfrac{s}{\sqrt{n}}$
Mean (σ known) — z-test	
Test stat	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
CI	$\bar{x} \pm z_{\alpha/2} \cdot \dfrac{\sigma}{\sqrt{n}}$
Proportion — One-proportion z-test	
Conditions	$np_0 \geq 10$ and $n(1-p_0) \geq 10$
Test stat	$z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$ (use p_0 in SE)
CI	$\hat{p} \pm z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$ (use $\hat{p}$ in SE)
Sample Size	
For mean	$n = \left(\dfrac{z_{\alpha/2} \cdot \sigma}{E}\right)^2$, E = margin of error
For prop	$n = \left(\dfrac{z_{\alpha/2}}{E}\right)^2 \hat{p}(1-\hat{p})$; use $\hat{p} = 0.5$ if unknown

TWO-SAMPLE INFERENCE (Lessons 23–25)

Independent Means — Two-sample t	
Conditions	Two indep. random samples; both normal or large n
Test stat	$t = \dfrac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$
CI	$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,\, df}\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$
df (conserv.)	$df = \min(n_1 - 1,\ n_2 - 1)$
Pooled t (assumes $\sigma_1^2 = \sigma_2^2$)	
s_p^2	$\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$
Test stat	$t = \dfrac{(\bar{x}_1 - \bar{x}_2) - D_0}{s_p\sqrt{1/n_1 + 1/n_2}}$, $df = n_1 + n_2 - 2$
Paired t (Lesson 24)	
Setup	$d_i = x_i - y_i$, then one-sample t on differences
Test stat	$t = \dfrac{\bar{d} - D_0}{s_d/\sqrt{n}} \sim t_{n-1}$
CI	$\bar{d} \pm t_{\alpha/2,\, n-1} \cdot \dfrac{s_d}{\sqrt{n}}$
When?	Same subjects measured twice; natural pairing
Two Proportions (Lesson 25)	
Conditions	$n_i\hat{p}_i \geq 10$ and $n_i(1-\hat{p}_i) \geq 10$ for $i = 1, 2$
Test stat	$z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}$ $\hat{p} = \dfrac{x_1 + x_2}{n_1 + n_2}$ (**pooled** under H_0)
CI	$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ (**unpooled** for CI)

POOLED vs. UNPOOLED vs. PAIRED

Pooled t	Assumes $\sigma_1^2 = \sigma_2^2$. Uses s_p^2 and $df = n_1 + n_2 - 2$. More powerful when assumption holds, not robust if violated.
Unpooled t	Default for indep. samples. No equal-variance assumption. Use $df = \min(n_1 - 1, n_2 - 1)$ for hand calculations.
Paired t	Use when observations are *naturally paired* (same subject, before/after, matched pairs). $df = n - 1$.

Note: If a pair-removal in Sample 1 forces a specific removal in Sample 2 ⇒ **paired**.

CONDITIONS CHECKLIST

All tests	Random sampling / random assignment
t-tests	Normal pop or $n \geq 30$ (CLT); indep. obs.
Paired t	Differences d_i approx. normal
Pooled t	Equal pop variances ($\sigma_1^2 = \sigma_2^2$)
z for prop	$np_0 \geq 10$, $n(1-p_0) \geq 10$ (one-prop) $n_i\hat{p}_i \geq 10$, $n_i(1-\hat{p}_i) \geq 10$ (two-prop)

REGRESSION, ANOVA, AND DECISION GUIDE

SIMPLE LINEAR REGRESSION (Lessons 28–30)

Model	$Y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$
Fitted line	$\hat{y} = b_0 + b_1 x$
Slope b_1	$\frac{S_{xy}}{S_{xx}} = r\frac{s_y}{s_x}, \quad S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y})$
Intercept b_0	$\bar{y} - b_1\bar{x}$
Residual	$e_i = y_i - \hat{y}_i$
SSE / SST	$\sum e_i^2 \ / \ \sum(y_i - \bar{y})^2$
R^2	$1 - \frac{\text{SSE}}{\text{SST}}$; for SLR, $R^2 = r^2$

Interpretation:

- **Slope:** avg. change in y per 1-unit increase in x.
- **Intercept:** predicted y when $x = 0$ (often not meaningful).
- **R^2:** proportion of variability in y explained by x.

Conditions (LINE): **L**inear relationship, **I**ndependent obs., **N**ormal residuals, **E**qual variance.
R: `lm(y ~ x)`; `summary(lm(y ~ x))`

MULTIPLE LINEAR REGRESSION (Lessons 31–32)

Model	$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$
Fitted	$\hat{y} = b_0 + b_1 x_1 + \cdots + b_k x_k$
Coef. test	$H_0 : \beta_j = 0 \quad t = \frac{b_j}{SE(b_j)} \sim t_{n-k-1}$
Adj. R^2	Penalizes adding predictors; preferred for model comparison

Coefficient interpretation: b_j is the avg. change in y per 1-unit increase in x_j, *holding all other predictors constant.*
Categorical predictors (Lesson 32):

- Convert to indicator (0/1) variables; one level becomes the **reference**.
- Coefficient = avg. difference between that level and the reference, holding other predictors constant.
- For a c-level categorical: $c - 1$ indicator variables.

Significance: predictor is significant at α if its individual p-value $< \alpha$. Failing to reject $\neq$ proving the coefficient is zero.
R: `lm(y ~ x1 + x2 + x3)`; `summary(...)` returns coefficient table with t and p.

ONE-WAY ANOVA (Lesson 36)

Setup: k groups, sample sizes $n_1, \ldots, n_k$, total $N = \sum n_i$.
Hypotheses: $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ vs. H_a : at least one μ_i differs.

Grand mean	$\bar{\bar{y}} = \frac{1}{N}\sum_{i,j} y_{ij}$
SSTr (Treatment)	$\sum_i n_i(\bar{y}_i - \bar{\bar{y}})^2, \quad df = k - 1$
SSE (Error)	$\sum_{i,j}(y_{ij} - \bar{y}_i)^2, \quad df = N - k$
SST (Total)	SSTr + SSE, $\quad df = N - 1$
MSTr	SSTr$/(k-1)$
MSE	SSE$/(N-k)$
F-statistic	$F = \frac{\text{MSTr}}{\text{MSE}} \sim F_{k-1,\,N-k}$
p-value	`1 - pf(F, k-1, N-k)` (always upper tail)

Conditions:

- Independent random samples within and between groups.
- Approximately normal populations within each group.
- Equal variances across groups: $\max(s_i)/\min(s_i) < 2$ (rule of thumb).

R: `aov(y ~ group)`; `summary(aov(...))`.

WHICH PROCEDURE DO I USE?

Data	Parameter	Procedure
1 quant sample	μ	One-sample t
1 categorical	p	One-proportion z
2 indep quant	$\mu_1 - \mu_2$	Two-sample t
2 paired quant	μ_d	Paired t
2 indep categ	$p_1 - p_2$	Two-proportion z
≥ 3 indep quant	$\mu_1, \ldots, \mu_k$	One-way ANOVA (F)
1 quant + 1 quant	relationship	Simple linear regression
1 quant + several quant/categ	relationship	Multiple linear regression

Key questions: (1) one or multiple groups? (2) quantitative or categorical response? (3) if two groups, are they independent or paired?

HYPOTHESIS TEST STEPS

1. **Hypotheses:** state H_0 and H_a in terms of population parameters.
2. **Conditions:** verify assumptions for the chosen test.
3. **Test statistic:** compute t, z, or F from data.
4. **p-value** (R-lite):
t-test: H_a :>: `1-pt(t,df)`; <: `pt(t,df)`; $\neq$: `2*(1-pt(abs(t),df))`
z-test: H_a :>: `1-pnorm(z)`; <: `pnorm(z)`; $\neq$: `2*(1-pnorm(abs(z)))`
F-test (ANOVA): `1 - pf(F, df1, df2)`
5. **Decision:** reject H_0 if $p < \alpha$; else FTR.
6. **Conclusion:** interpret in context.

INTERPRETATION TEMPLATES

Confidence Interval:
"We are [CL]% confident that the true [parameter in context] is between [lower] and [upper]."
HT Conclusion:
"With a p-value of _____ which is less than α, we reject H_0, meaning that [context of H_a]."
"With a p-value of _____ which is greater than α, we FTR H_0, meaning that [insufficient evidence for context of H_a]."
Regression coefficient:
"Holding [other predictors] constant, each 1-unit increase in [x_j] is associated with an avg. [increase/decrease] of [b_j] in [y]."
R^2: "[$R^2 \times 100$]% of the variability in [y in context] is explained by [the predictors in context]."

COMMON MISTAKES TO AVOID

× "95% probability the true mean is in this interval"
✓ "95% of such intervals contain the true mean"
× "Accept H_0" ✓ "Fail to reject H_0"
(absence of evidence $\neq$ evidence of absence)
× Stating hypotheses about $\bar{x}$, $\hat{p}$ ✓ State about μ, p
× Confusing $P(A|B)$ with $P(B|A)$ in Bayes'
× Interpreting MLR coefficient without "holding others constant"
× "Failed to reject" ⇒ "coefficient is zero"
(FTR means insufficient evidence; effect may still exist)
× Statistical significance ⇒ practical importance
(small p does NOT imply large or meaningful effect)

COMMON CRITICAL VALUES (R-LITE)

CL	$z_{\alpha/2}$	R-lite
90%	1.645	`qnorm(0.95)`
95%	1.960	`qnorm(0.975)`
99%	2.576	`qnorm(0.995)`

t **critical:** `qt(1 - alpha/2, df)`
F **critical:** `qf(1 - alpha, df1, df2)`

General Reminders: State hypotheses about *population parameters*, not statistics • Verify conditions before testing • Use R-lite (`pdist`, `qdist`) for p-values and critical values • Report conclusions in context • CIs and HTs are complementary — use both when asked • Statistical significance $\neq$ practical significance • For regression and ANOVA: significance of a predictor / overall test $\neq$ practical importance of magnitude