STAT115 Content Speed-Run

Self-Study Pack (Active Recall + Micro Practice + R Mini-Kit)

Prepared for catch-up after a six-week absence

How to use this pack (learning-science built-in)

- **Active recall first, reread last:** answer from memory before checking. Speaking your answer out loud improves retention.
- **Dual coding:** sketch tiny diagrams (axes, curves, residual plots) alongside formulas and code.
- **Spacing & interleaving:** tick the review boxes (Day 0/2/7/14) and mix topics during review.
- **Error log:** when you miss a recall item, write why and how you will avoid it next time.

Contents

Lecture 1 — Orientation & What Statistics Is

Core Content (2–6 min skim)

- Statistics is **learning from data** and about **describing and quantifying variability**.
- Tutorials are highly recommended; R is available on lab machines.
- Final exam: **3 hours**, about **90 multiple-choice** questions. Final grade: $F = 0.7 \times Exam + 0.3 \times Assignments$.

Active Recall (cover *Core Content* above; answer from memory)

1. Complete: "Statistics is ____________________."
2. Besides learning from data, what two words describe the focus of statistics?
3. What is the final-exam format and duration?
4. How is the final mark calculated?
5. One reason tutorials add value?

Micro Practice (5–10 min)

Find two tutorial slots you can attend and write them here. Commit on paper as well.

R Mini-Kit (copy & run)

No code yet—just ensure you can open RStudio and run: `1 + 1`.

Spaced Review: ☐ Day 0 ☐ Day 2 ☐ Day 7 ☐ Day 14

Lecture 2 — Statistical Software (R focus)

Core Content (2–6 min skim)

- We will use **R** via RStudio. Excel is common but limited for robust statistical analysis.
- Minimal R toolkit suffices: read data, summarise, tabulate, test, model, diagnose.

Active Recall (cover *Core Content* above; answer from memory)

1. Why is R preferred over pure spreadsheets for analysis?
2. What does RStudio add on top of base R?
3. Name two other statistical packages you know.

Micro Practice (5–10 min)

Create a new R script. Type the commands below and run them without errors.

R Mini-Kit (copy & run)

```
# Reading and peeking at data
D <- read.csv("yourfile.csv")
head(D); summary(D)

# Categorical tabulation
T <- table(D$A, D$B); T
prop.table(T)       # overall proportions
prop.table(T, 1)    # row proportions
prop.table(T, 2)    # column proportions
```

Spaced Review: ☐ Day 0 ☐ Day 2 ☐ Day 7 ☐ Day 14

Lecture 3 — Contingency Tables & Basic Probability

Core Content (2–6 min skim)

- Contingency tables show **counts** and **proportions**; treat proportions as probabilities for practice.
- **Marginal** probabilities are in the margins; **joint** inside cells; **conditional** restrict to a row/column.
- Independence fails if $\Pr(Survival \mid Sex) \neq \Pr(Survival)$.

Active Recall (cover *Core Content* above; answer from memory)

1. Where do marginal probabilities live?
2. If total $= 2092$ and female-survivors $= 316$, compute $\Pr(\text{female} \wedge \text{survived})$.
3. Explain in words why survival and sex are not independent in Titanic data.
4. How do you convert a count table to proportions?
5. Define "joint" vs "conditional" probability in one sentence each.

Micro Practice (5–10 min)

Using Titanic counts, calculate $\Pr(S), \Pr(M), \Pr(S \wedge M), \Pr(S \mid M)$, then check independence.

R Mini-Kit (copy & run)

```
# titanic: 2x2 table of counts, rows=sex, cols=survival
Total <- sum(titanic)
P <- titanic / Total; P
# Marginals
Pr_S <- margin.table(P, 2)["yes"]
Pr_M <- margin.table(P, 1)["male"]
# Conditional
Pr_S_given_M <- P["male","yes"] / Pr_M
```

Spaced Review: ☐ Day 0 ☐ Day 2 ☐ Day 7 ☐ Day 14

Lecture 4 — Populations, Parameters, Normal Model (First Look)

Core Content (2–6 min skim)

- **Population vs sample**; **parameter** (μ, σ) vs **statistic** $(\bar{y}, s)$.
- Estimation targets parameters; the **Normal** distribution often models quantitative data.

Active Recall (cover *Core Content* above; answer from memory)

1. Give one parameter and its sample-statistic counterpart.
2. Why introduce a distributional model like the Normal?
3. What do $\bar{y}$ and s estimate?

Micro Practice (5–10 min)

Sketch a bell curve; mark μ and $\pm 2\sigma$. Write what "about 95%" means under Normal.

R Mini-Kit (copy & run)

```
x <- rnorm(100, mean=0, sd=1)
mean(x); sd(x)
hist(x)  # quick visual
```

Spaced Review: ☐ Day 0 ☐ Day 2 ☐ Day 7 ☐ Day 14

Lecture 5 — Confidence Intervals (CIs), Confidence Level, SE, Sample Size

Core Content (2–6 min skim)

- `t.test()` yields CIs for a mean; increasing `conf.level` widens the CI.
- Standard error of $\bar{y}$: $s/\sqrt{n}$. Larger s widens; larger n narrows (all else fixed).
- Design question: choose n to hit a target margin of error (MOE).

Active Recall (cover *Core Content* above; answer from memory)

1. How does raising `conf.level` affect CI width?
2. Write SE($\bar{y}$).
3. Two levers to narrow a CI?
4. Plain-English meaning of a 95% CI?
5. Why is it unethical to overstate n?

Micro Practice (5–10 min)

Run `t.test(GAG$conc, conf.level = 0.90/0.95/0.99)`. Which is widest? Why?

R Mini-Kit (copy & run)

```
out95 <- t.test(GAG$conc, conf.level = 0.95)
out99 <- t.test(GAG$conc, conf.level = 0.99)
out90 <- t.test(GAG$conc, conf.level = 0.90)
# Sample-size sketch for MOE (xi) using a pilot s
z <- qnorm(1-0.05/2); s <- sd(GAG$conc); xi <- 0.04
n_needed <- ceiling((z*s/xi)^2)
```

Spaced Review: ☐ Day 0 ☐ Day 2 ☐ Day 7 ☐ Day 14

Lecture 6 — Two Independent Means (Welch Two-Sample t)

Core Content (2–6 min skim)

- Use `t.test(x, y)` for **independent groups** (Welch by default): outputs t, df, p, CI, and group means.
- Interpretation: p-value measures **incompatibility with** H_0; **CI** indicates plausible effect size.
- With small samples, normality matters more; be cautious.

Active Recall (cover *Core Content* above; answer from memory)

1. State H_0 and H_A for comparing two means.
2. What does Welch guard against vs pooled-variance t?
3. Why doesn't the p-value tell "how big" the effect is?
4. Which parameter does the CI estimate here (write $\mu_1 - \mu_2$)?
5. One assumption to check in each group?

Micro Practice (5–10 min)

Given `control$Freq` and `solitary$Freq`, run `t.test(control$Freq, solitary$Freq)` and interpret: Is 0 inside the CI? Which group mean is higher and by how much (roughly)?

R Mini-Kit (copy & run)

```
out <- t.test(control$Freq, solitary$Freq)
out$estimate      # group means (mind the order)
out$conf.int      # CI for mu_control - mu_solitary
out$p.value
```

Spaced Review: ☐ Day 0 ☐ Day 2 ☐ Day 7 ☐ Day 14

Lecture 7 — Paired Data (Within-Subject)

Core Content (2–6 min skim)

- Paired design: each observation in A corresponds to one in B; analyze **differences**.
- Two equivalent paths: (1) compute differences and one-sample t; (2) `t.test(A,B, paired=TRUE)`.
- The CI from both approaches is **identical**; wording differs.

Active Recall (cover *Core Content* above; answer from memory)

1. Why analyze paired data via differences?
2. What parameter is tested in paired t (write μ_d)?
3. How do the two outputs differ in wording but not numbers?
4. Give a real-world example that should be analyzed as paired.
5. What goes wrong if you treat paired observations as independent?

Micro Practice (5–10 min)

For auditory/visual reaction times, create a difference variable and run both analyses. Confirm the same CI.

R Mini-Kit (copy & run)

```
AV <- read.csv("AV.csv")
AV$differ <- AV$visual - AV$auditory
# Option 1
one <- t.test(AV$differ)
# Option 2 (equivalent CI)
two <- t.test(AV$visual, AV$auditory, paired=TRUE)
one$conf.int; two$conf.int
```

Spaced Review: ☐ Day 0 ☐ Day 2 ☐ Day 7 ☐ Day 14

Lecture 8 — Simple Linear Regression (fit → diagnose)

Core Content (2–6 min skim)

- Model: $y = \beta_0 + \beta_1 x + \varepsilon$. Fitted by least squares (minimise squared residuals).
- Interpret β_1 as expected change in y per 1-unit increase in x (when sensible).
- Be cautious interpreting β_0 if $x = 0$ lies outside observed range.

Active Recall (cover *Core Content* above; answer from memory)

1. In words, what are fitted values and residuals?
2. Explain β_1 in your own words.
3. Why might β_0 be uninterpretable in some data sets?

Micro Practice (5–10 min)

Fit a line predicting possum head length from total length. Write one sentence interpreting β_1.

R Mini-Kit (copy & run)

```
m <- lm(head_l ~ total_l, data=possum)
coef(m); fitted(m); residuals(m)
```

Spaced Review: ☐ Day 0 ☐ Day 2 ☐ Day 7 ☐ Day 14

Lecture 9 — Checking SLR Assumptions (LINE) with Residuals

Core Content (2–6 min skim)

- Assumptions: **LINE** = Linearity, Independence, Normality, Equal variance of errors.
- Use **studentised residuals** and **residuals vs fitted** to diagnose trend (linearity), funnel (variance), outliers.

Active Recall (cover *Core Content* above; answer from memory)

1. Expand LINE.
2. Which plot do you look at first to check assumptions?
3. High-level meaning of a studentised residual?
4. Name one worrying pattern in residuals-vs-fitted.
5. Why check assumptions after fitting?

Micro Practice (5–10 min)

Make the residual plot for the possum model; add a horizontal line at 0. Note any trends or funnels.

R Mini-Kit (copy & run)

```
fit <- lm(head_l ~ total_l, data=possum)
rvf <- rstudent(fit)
plot(fitted(fit), rvf); abline(h=0)
```

Spaced Review: ☐ Day 0 ☐ Day 2 ☐ Day 7 ☐ Day 14

— Fast Review Sheet (pin on your wall)

- **Contingency tables** → marginal, joint, conditional; independence check: $\Pr(A \mid B) \stackrel{?}{=} \Pr(A)$.
- **CIs:** width ↑ with conf.level ↑ or s ↑; width ↓ with n ↑.
- **Welch two-sample** t for independent groups; **paired** t for within-subject differences.
- **SLR:** fit with `lm(y ~ x)`; check **LINE** via residual plots and studentised residuals.