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

Lecture 10 Sampling Distributions and Central Limit Theorem (CLT)

Core Content (2–6 min skim)

- A sampling distribution is the distribution of a statistic (for example, $\bar{y}$) under repeated sampling.
- CLT (informal): for large n, the sampling distribution of $\bar{y}$ is approximately Normal with mean μ and SE $s/\sqrt{n}$, regardless of the population shape (mild conditions).
- Standard error connects sample size and variability: larger n leads to smaller SE and tighter CIs.

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

1. Define sampling distribution in one sentence.
2. State the CLT informally for the sample mean.
3. How does SE($\bar{y}$) scale with n?

Micro Practice (5–10 min)

Simulate 2000 sample means from a skewed distribution (for example, Exponential). Make a histogram and compare with a Normal curve.

R Mini-Kit (copy & run)

```
set.seed(1)
means <- replicate(2000, mean(rexp(50, rate = 1)))
hist(means, breaks = 30, main = "Sampling distribution of ybar")
mean(means); sd(means)
```

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

Lecture 11 Proportions and the Binomial Model

Core Content (2–6 min skim)

- Binary outcomes (success or failure) can be modelled with Binomial(n, p).
- Sample proportion $\hat{p} = X/n$ estimates p; $\text{SE}(\hat{p}) \approx \sqrt{\hat{p}(1-\hat{p})/n}$.
- Normal approximation works when np and $n(1-p)$ are not too small.

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

1. Define $\hat{p}$ and its approximate SE.
2. When is the Normal approximation to the Binomial reasonable?
3. Give a real example where a proportion is the right summary.

Micro Practice (5–10 min)

From 250 patients, 37 show a side effect. Compute $\hat{p}$ and an approximate 95% CI.

R Mini-Kit (copy & run)

```
x <- 37; n <- 250
phat <- x/n
se <- sqrt(phat*(1-phat)/n)
ci <- phat + c(-1,1)*qnorm(0.975)*se
phat; ci
```

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

Lecture 12 One-Sample Proportion: Confidence Interval and Test

Core Content (2–6 min skim)

- CI for p: `prop.test(x, n)` gives an approximate (Wilson-like) CI; `binom.test` gives an exact CI.
- Hypothesis test for p with null p_0: z (approximate) or exact binomial test.
- Report both $\hat{p}$ and the CI; p-value addresses compatibility with H_0.

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

1. Which R function gives an exact binomial test and CI?
2. In words, what does a small p-value tell you about H_0?
3. Why can an exact method be preferable with small n?

Micro Practice (5–10 min)

Test whether the true adverse-event rate differs from 10% when $x = 37, n = 250$.

R Mini-Kit (copy & run)

```
binom.test(37, 250, p = 0.10)
prop.test(37, 250, p = 0.10, correct = TRUE)
```

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

Lecture 13 Two Proportions: Difference, CI, and Test

Core Content (2–6 min skim)

- Compare $p_1 - p_2$ with a two-sample test for proportions; use `prop.test(x = c(x1,x2), n = c(n1,n2))`.
- Report the CI for $p_1 - p_2$ and interpret direction and magnitude.
- Avoid over-interpreting p-values without effect-size context.

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

1. Write the parameter of interest for two proportions.
2. Name one assumption that justifies the Normal approximation here.
3. What does it mean if 0 is inside the CI for $p_1 - p_2$?

Micro Practice (5–10 min)

Group A: 18/120 successes; Group B: 29/150. Test for a difference and give the 95% CI.

R Mini-Kit (copy & run)

```
prop.test(x = c(18,29), n = c(120,150))
```

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

Lecture 14 Association in 2x2 Tables: Risk Difference, Risk Ratio, Odds Ratio

Core Content (2–6 min skim)

- Risk difference (RD): $p_1 - p_2$. Risk ratio (RR): p_1/p_2. Odds ratio (OR): $\frac{p_1/(1-p_1)}{p_2/(1-p_2)}$.
- Interpretation depends on design (cohort vs case-control). OR approximates RR when outcomes are rare.
- Always report context and time frame.

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

1. Define RD, RR, and OR.
2. When is OR approximately equal to RR?
3. Give one pitfall when interpreting ratios.

Micro Practice (5–10 min)

From a 2x2 table, compute RD, RR, and OR. Which is most interpretable for patients in your context?

R Mini-Kit (copy & run)

```
# Suppose tab is matrix(c(a,b,c,d), nrow=2, byrow=TRUE)
prop <- prop.table(tab, 1)
p1 <- prop[1,2]; p2 <- prop[2,2]
RD <- p1 - p2
RR <- p1 / p2
OR <- (p1/(1-p1)) / (p2/(1-p2))
c(RD=RD, RR=RR, OR=OR)
```

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

Lecture 15 Study Design: Experiments vs Observational, Bias and Confounding

Core Content (2–6 min skim)

- Randomisation, control, and blinding reduce bias; observational studies are vulnerable to confounding.
- Always state the unit of analysis, sampling frame, and inclusion or exclusion criteria.
- Association is not causation; consider DAG-like thinking to name potential confounders.

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

1. One difference between experimental and observational designs.
2. Define confounding in one sentence.
3. Name two common sources of bias.

Micro Practice (5–10 min)

Take a claim from news or social media. Identify whether the underlying study is experimental or observational and list likely confounders.

R Mini-Kit (copy & run)

```
# Simple random sample indices
i <- sample.int(nrow(D), size = 100)
D_s <- D[i,]
```

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

Lecture 16 Errors, Power, and Planning

Core Content (2–6 min skim)

- Type I error (false positive) rate is set by α; Type II error relates to $1-$power.
- Power increases with larger effects, larger n, smaller variability, and higher α (trade-offs apply).
- Pre-specify hypotheses and primary outcomes to avoid p-hacking.

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

1. Define power in words.
2. Name two knobs that increase power (holding others fixed).
3. Why is multiple testing dangerous without correction?

Micro Practice (5–10 min)

Sketch how the required n changes when the target effect size halves (qualitatively). What happens to power if n stays fixed?

R Mini-Kit (copy & run)

```
# Crude simulation of power for a one-sample t under mean shift
delta <- 0.3; n <- 40; B <- 1000
pvals <- replicate(B, t.test(rnorm(n, mean=delta, sd=1))$p.value)
mean(pvals < 0.05)   # approximate power
```

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

Lecture 17 Correlation (Pearson r) and Scatterplots

Core Content (2–6 min skim)

- Pearson r measures linear association (from -1 to 1); it is unitless and symmetric in x and y.
- Nonlinear patterns can yield r near 0 even when variables are strongly related.
- Outliers can distort r; always inspect the scatterplot.

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

1. What does the sign and magnitude of r indicate?
2. Why must you always look at the scatterplot before trusting r?
3. Give one situation where r is inappropriate.

Micro Practice (5–10 min)

Compute r between two quantitative variables and draw the scatterplot. Describe form, strength, and outliers.

R Mini-Kit (copy & run)

```
plot(D$x, D$y)
cor(D$x, D$y)
cor.test(D$x, D$y)
```

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

Lecture 18 Transformations and Nonlinearity

Core Content (2–6 min skim)

- Log or square-root transforms can stabilise variance and linearise relationships.
- Interpret coefficients on the transformed scale carefully (for example, log-y implies multiplicative effects).
- Compare residual plots before and after transformation.

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

1. Name one reason to take logs of y.
2. After logging y, how would you interpret a slope of 0.07?
3. What visual cue in residuals-versus-fitted suggests a variance problem?

Micro Practice (5–10 min)

Fit models with y and with log(y); compare residual diagnostics and R output.

R Mini-Kit (copy & run)

```
m1 <- lm(y ~ x, data=D)
m2 <- lm(log(y) ~ x, data=D)
par(mfrow=c(1,2))
plot(fitted(m1), rstudent(m1)); abline(h=0)
plot(fitted(m2), rstudent(m2)); abline(h=0)
par(mfrow=c(1,1))
```

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

Lecture 19 Outliers, Leverage, and Influence

Core Content (2–6 min skim)

- Leverage points have unusual x; influential points change fit noticeably (for example, large Cook distance).
- Check hat values and Cook distance; diagnose, then justify any exclusions transparently.
- Refit with and without suspicious points and compare conclusions.

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

1. Distinguish leverage and influence.
2. Name two diagnostics for influence.
3. Why is pre-specifying exclusion rules important?

Micro Practice (5–10 min)

Identify the top three most influential observations in your regression and inspect their raw records.

R Mini-Kit (copy & run)

```
fit <- lm(y ~ x, data=D)
h <- hatvalues(fit)
cd <- cooks.distance(fit)
head(sort(cd, decreasing=TRUE), 3)
```

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

Lecture 20 Prediction and Intervals in Regression

Core Content (2–6 min skim)

- Use `predict(..., interval="confidence")` for mean response; use `interval="prediction"` for a new individual.
- Prediction intervals are wider than confidence intervals.
- Do not extrapolate far beyond observed x.

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

1. Difference between a confidence interval for the mean response and a prediction interval.
2. Why are prediction intervals wider?
3. What is extrapolation and why is it risky?

Micro Practice (5–10 min)

Fit an SLR and compute both CI and PI at x values near the center of your data. Compare widths.

R Mini-Kit (copy & run)

```
fit <- lm(y ~ x, data=D)
new <- data.frame(x = c(10, 20))
conf <- predict(fit, new, interval = "confidence")
pred <- predict(fit, new, interval = "prediction")
conf; pred
```

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

Lecture 21 ANOVA (Concept) and Categorical Predictors

Core Content (2–6 min skim)

- One-way ANOVA tests equality of k group means; equivalent to regression with k-1 dummy variables.
- Assumptions mirror SLR errors: independence, Normality within groups, equal variances.
- Report group means with CIs and the overall F test; follow up with planned contrasts where relevant.

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

1. What does one-way ANOVA test?
2. Name the error assumptions for ANOVA.
3. How do you represent a 4-level factor in regression?

Micro Practice (5–10 min)

Run one-way ANOVA for y across a 3-level factor and show group means with 95% CIs.

R Mini-Kit (copy & run)

```
fit <- aov(y ~ group, data=D)
summary(fit)
aggregate(y ~ group, data=D, FUN=mean)
T <- TukeyHSD(fit)
T
```

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

Fast Review Sheet (pin on your wall)

- **Probability and tables:** marginal / joint / conditional; independence check: $\Pr(A \mid B) \stackrel{?}{=} \Pr(A)$. For 2x2, also report **RD**, **RR**, **OR** (when events are rare, OR $\approx$ RR).
- **Sampling and CLT:** the sampling distribution of $\bar{y}$ is approximately Normal for large n; $\mathrm{SE}(\bar{y}) = s/\sqrt{n}$.
- **CIs and SEs:** width increases with conf.level or s; width decreases with n. For a proportion $\hat{p}$: $\mathrm{SE}(\hat{p}) \approx \sqrt{\hat{p}(1-\hat{p})/n}$.
- **Means:** one-sample t for a mean; **Welch two-sample** t for independent groups; **paired** t on differences for within-subject designs. Always report estimate, CI, and a one-sentence practical interpretation.
- **SLR:** model $y = \beta_0 + \beta_1 x + \varepsilon$. Interpret β_1 as change in y per 1-unit increase in x; β_0 may be non-sensical if $x = 0$ is outside the data. Check **LINE**; inspect residuals, leverage, and Cook distance.
- **Prediction vs confidence:** `predict(..., interval="confidence")` is for the mean response; `interval="prediction"` is for a new individual (wider).
- **Power and planning:** power increases with larger effects, larger n, smaller variability, and higher α (trade-offs). Pre-specify outcomes; avoid p-hacking.
- **Minimal R verbs to remember:** `t.test(...)`; `prop.test(...)` / `binom.test(...)`; `table` and `prop.table`; `lm(y~x)`; `plot(fitted(), rstudent())`; `predict(..., interval=)`.

How to schedule your catch-up (suggested)

- **Week 1 (catch-up):** two lectures per day (about 35–45 min each).
  - For each lecture: 6 min Core skim → 5 min Active Recall (eyes off) → 8–10 min Micro Practice → 5–8 min R Mini-Kit.
  - End of day: 15 min mixed recall (pick 6–8 questions across the day's lectures).
- **Week 2 (consolidate):** revisit each lecture on Day 2 and Day 7; do only Active Recall plus one Micro Practice. Tick the review boxes: Day 0 / 2 / 7 / 14.
- **Error log protocol:** when you miss an item, write the reason and a fix in a one-page sheet you see daily.
- **R reps:** once per day, retype one Mini-Kit from memory (no copy-paste) to keep commands fluent.

- **Exam warm-up:** 3 blocks $\times$ 30 MCQs under time; after each block, classify errors (concept vs slip) and fix with one targeted Micro Practice.