STATISTICS · STUDY GUIDE

Exam 7 Study Guide

This guide covers chapters 13 and 14, the last two tools in the course. You will test categorical data against a claim and for association with chi-square, compare three or more means with one-way ANOVA, check the conditions and find which groups differ. Then you will describe two quantitative variables with a scatterplot and r, fit the least-squares line, read residuals and r^2, and predict honestly. Work one chapter at a time, try each worked example with the steps covered, and finish with the ten self-test items. The cram sheet at the end holds every formula.

WHAT THIS GUIDE COVERS

- Chapter 13 · chi-square goodness of fit and test of independence, expected counts and df, standardized residuals and Cramér's V, one-way ANOVA and the F ratio, conditions for both, and post-hoc comparisons with Tukey and Bonferroni.
- Chapter 14 · scatterplots, the correlation r and its limits, the least-squares line and its interpretation, residuals, r^2 and residual plots, the standard error of the estimate, and prediction without extrapolation.

CHAPTER 13

Chi-Square and ANOVA

Read it in full: **Chapter 13 in the textbook**. Practice it in the lab: **Test of Independence** and **Between and Within**.

What to know

The **goodness-of-fit** test compares one categorical variable's observed counts with a claimed distribution. H0: the proportions are exactly as claimed; Ha: at least one differs. Expected counts are $E = n \times p$; do not round them, and check they sum to n. The statistic $\chi^2 = \Sigma(O - E)^2 \div E$ is never negative, and only big values count against H0, so the test is always **right-tailed**. $df = k - 1$ (minus one more for each parameter estimated from the data).

In the book's meal-plan example (claimed 25/40/35, observed 40, 95, 65 of 200), $\chi^2 = 2.0000 + 2.8125 + 0.3571 = 5.17$ with df = 2. The critical value at $\alpha = 0.05$ is 5.991, $p \approx 0.075$, so we fail to reject. That does not make the claim true. The biggest term, Silver at 2.81, shows where to look next.

The **test of independence** crosses two categorical variables in an r × c table. H0: no association; Ha: associated. Expected counts are E = (row total × column total) ÷ grand total, the counts you would see if every row had the same column profile. df = (r − 1)(c − 1). The book's exercise and sleep table gives $\chi^2 = 16.07$ on 2 df, $p \approx 0.0003$. Describe the pattern with **standardized residuals** (O − E) ÷ √E (beyond about ±2 marks a driving cell) and the size with **Cramér's V** = √[χ^2 ÷ (n × min(r − 1, c − 1))], here 0.23. Association is not causation.

Running many t-tests inflates the **family-wise error rate**: with three groups, $1 - 0.95^3 \approx 14\%$, and with five groups (ten tests) about 40%, as rough illustrations. **One-way ANOVA** tests H0: $\mu_1 = \mu_2 = ... = \mu_k$ against "at least one mean differs" (not "all differ") with one test and one α.

ANOVA compares two estimates of σ^2. **MSW** pools the scatter inside the groups (noise); **MSB** measures how far group means sit from the grand mean (signal plus noise). F = MSB ÷ MSW sits near 1 when the means are equal and grows when they are not, so the test is right-tailed. $SS_{between} = \Sigma n_j(\bar{x}_j - \bar{x})^2$ on k − 1 df; $SS_{within} = \Sigma\Sigma(x - \bar{x}_j)^2$ on N − k df; $SS_{total} = SS_{between} + SS_{within}$ on N − 1 df, a built-in check. The book's three methods give F = 125.00 ÷ 11.50 = 10.87 on (2, 12), $p \approx 0.0020$, and $\eta^2 = 250 \div 388 = 0.64$. With two groups, $F = t^2$ from the pooled t-test.

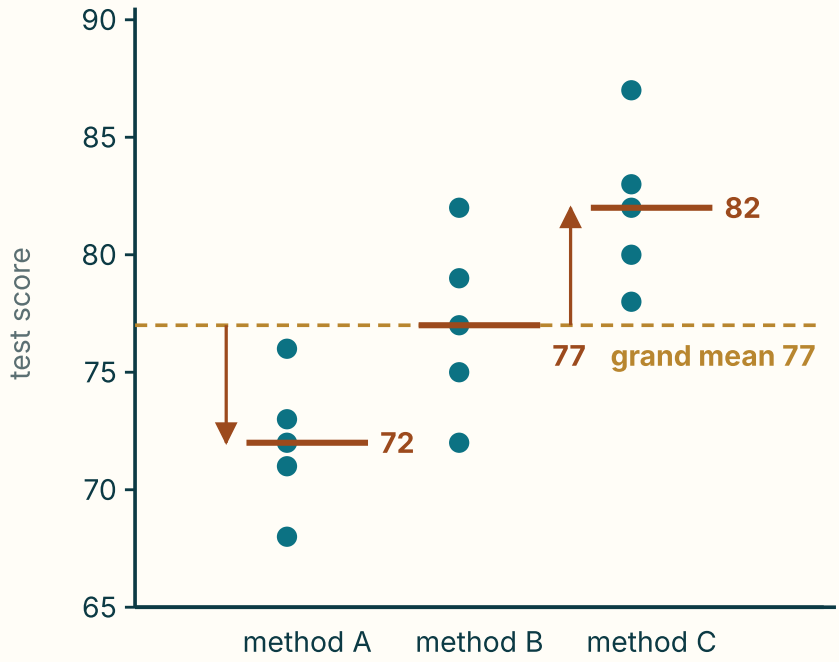
Chi-square needs **counts** (not percents), independent observations each in one cell, a random sample, and all **expected** counts at least 5 (Cochran allows up to 20% between 1 and 5, none below 1). If counts are too small, combine categories or use Fisher's exact test; paired categorical data need McNemar's test.

ANOVA needs independence within and between groups (the least forgiving), approximate normality within groups (robust with similar group sizes and decent n), **equal variances** (largest s under about twice the smallest, or largest variance under four times the smallest) and a quantitative response. Fixes: Welch's ANOVA for unequal variances, Kruskal-Wallis for strong

non-normality in small groups, repeated-measures ANOVA when the same subjects are in every group.

A significant F says only that at least one mean differs. **Post-hoc** tests find which pairs while holding the family-wise error at α. **Bonferroni** tests each of m comparisons at α ÷ m. **Tukey's HSD** = q × √(MSW ÷ n) is the standard for all pairs with equal n. In the book, HSD = 3.77 × 1.517 = 5.72, so only A versus C (10 points apart) differs; A versus B and B versus C (5 apart) do not. If F is not significant, do not go fishing through the pairs.

> **LOCK THESE IN**
>
> - $\chi^2 = \Sigma(O - E)^2/E$; right-tailed; never negative.
> - Goodness of fit: E = np, df = k − 1. Never round E.
> - Independence: E = (row × column)/total, df = (r − 1)(c − 1).
> - Chi-square needs counts and every expected count ≥ 5.
> - F = MSB/MSW; df = k − 1 and N − k; $SS_T = SS_B + SS_W$.
> - Ha for ANOVA: at least one mean differs, not all.
> - Equal variances: largest s < 2 × smallest.
> - Post-hoc: Tukey HSD = q√(MSW/n); Bonferroni tests at α/m.



between (signal)
SS = 250, df = 2
MSB = 125.00

within (noise)
SS = 138, df = 12
MSW = 11.50

F = 125.00 ÷ 11.50 = 10.87

Arrows: group means sit far from 77 compared with the scatter inside each group.

F compares signal with noise. The three method means (72, 77, 82) spread far from the grand mean compared with the scatter inside each group, so F = 10.87 on (2, 12) df.

Formulas and when to use them

Quantity	Formula	When to use it
Chi-square statistic	$\chi^2 = \Sigma(O - E)^2 \div E$	Both chi-square tests
Expected, goodness of fit	$E = n \times p_i$	One variable versus a claim
Expected, independence	E = (row total × column total) ÷ grand total	Two-way table
df	k − 1; (r − 1)(c − 1)	GOF; independence
Standardized residual	$(O - E) \div \sqrt{E}$	Which cells drive the result
Cramér's V	$\sqrt{[\chi^2 \div (n \times \min(r - 1, c - 1))]}$	Size of an association, 0 to 1
Between SS	$\Sigma\, n_j(\bar{x}_j - \bar{x})^2$, df = k − 1	ANOVA signal
Within SS	$\Sigma\Sigma(x - \bar{x}_j)^2$, df = N − k	ANOVA noise
F	MSB ÷ MSW, MS = SS ÷ df	Equal means?
Eta squared	$\eta^2 = SS_{between} \div SS_{total}$	Share of variation explained
Tukey HSD	$q_{\alpha,\, k,\, N-k} \times \sqrt{(MSW \div n)}$	All pairs, equal n
Bonferroni	Each test at $\alpha \div m$	A few planned comparisons

Critical values at α = 0.05

Distribution	df	Critical value
χ^2	1, 2, 3, 4, 5	3.841, 5.991, 7.815, 9.488, 11.070
χ^2	6	12.592
F	(2, 12)	3.89
F	(2, 20) and (3, 20)	3.49 and 3.10
F	(3, 30) and (4, 30)	2.92 and 2.69
Studentized range q	k = 3, df = 12	3.77

Worked examples

WORKED EXAMPLE 1 · GOODNESS OF FIT FOR A DIE

A die is rolled 120 times: faces 1 to 6 come up 15, 22, 18, 26, 17, 22 times. Is it fair? Use α = 0.05.

STEP 1 · HYPOTHESES AND E

H0: each face has probability 1/6. E = 120 × 1/6 = 20 for every face, summing to 120. All E ≥ 5.

STEP 2 · TERMS

$(15 - 20)^2/20 = 1.25$; $(22 - 20)^2/20 = 0.20$; $(18 - 20)^2/20 = 0.20$; $(26 - 20)^2/20 = 1.80$; $(17 - 20)^2/20 = 0.45$; $(22 - 20)^2/20 = 0.20$.

STEP 3 · X^2

Sum = **4.10**, df = 6 − 1 = 5.

STEP 4 · DECIDE

Critical value 11.070; p ≈ 0.535. 4.10 < 11.070, so **fail to reject**.

STEP 5 · SAY IT

These rolls give no convincing evidence that the die is unfair. That is not proof of fairness; 120 rolls cannot detect a small bias. Face 4, with the largest term, is the one to watch.

WORKED EXAMPLE 2 · A TEST OF INDEPENDENCE WITH RESIDUALS AND V

240 commuters are classified by age (under 35: car 40, transit 35, active 25; 35 and over: car 90, transit 30, active 20). Test for association at $\alpha = 0.05$.

STEP 1 · MARGINS

Rows 100 and 140; columns 130, 65, 45; total 240.

STEP 2 · EXPECTED

Under 35: 54.17, 27.08, 18.75. 35 and over: 75.83, 37.92, 26.25. All at least 5.

STEP 3 · X^2

3.705 + 2.314 + 2.083 + 2.647 + 1.653 + 1.488 = **13.89**, df = (2 − 1)(3 − 1) = 2.

STEP 4 · DECIDE

13.89 > 5.991; p ≈ 0.0010. **Reject H0**: commute mode and age group are associated.

STEP 5 · DESCRIBE

Standardized residuals, under 35: −1.92, 1.52, 1.44; 35 and over: 1.63, −1.29, −1.22. The car cells have the largest residuals, though none passes ±2: younger commuters drive less than independence predicts and older ones more, with the transit and active cells reversed.

STEP 6 · SIZE

$V = \sqrt{(13.89 \div (240 \times 1))} =$ **0.24**, a modest association. These are survey data, so no causal claim.

WORKED EXAMPLE 3 · A ONE-WAY ANOVA BY HAND

Three study apps, four students each: A 10, 12, 11, 13; B 14, 15, 13, 16; C 12, 13, 14, 13. Test for a difference in means at $\alpha = 0.05$.

STEP 1 · MEANS

$\bar{x}_A = 11.5$, $\bar{x}_B = 14.5$, $\bar{x}_C = 13.0$; grand mean 13.00.

STEP 2 · BETWEEN

$4(-1.50)^2 + 4(1.50)^2 + 4(0.00)^2 =$ **18.00**, df = 2.

STEP 3 · WITHIN

5 + 5 + 2 = **12**, df = 12 − 3 = 9. Check: $SS_{total} = 30.00 = 18.00 + 12$.

STEP 4 · F

MSB = 18.00 ÷ 2 = 9.00; MSW = 12 ÷ 9 = 1.333; F = **6.75**.

STEP 5 · DECIDE

Critical F(2, 9) = 4.26; p ≈ 0.0162. Reject H0: at least one mean differs. $\eta^2 = 18.00 \div 30.00 = 0.60$.

STEP 6 · CHECK VARIANCES

Group SDs 1.29, 1.29, 0.82: largest ÷ smallest = 1.58, under 2.

WORKED EXAMPLE 4 · WHICH PAIRS DIFFER? TUKEY AND BONFERRONI

Continue the study-app ANOVA. Find which pairs of means differ.

STEP 1 · HSD

q (α = 0.05, k = 3, df = 9) = 3.95. √(MSW ÷ n) = √(1.333 ÷ 4) = 0.577. HSD = **2.28**.

STEP 2 · PAIRS

A vs B: 3.0; A vs C: 1.5; B vs C: 1.5.

STEP 3 · COMPARE

A vs B: significant; A vs C: not significant; B vs C: not significant.

STEP 4 · BONFERRONI CHECK

SE of a difference = √[1.333(1/4 + 1/4)] = 0.816. A vs B: t = 3.67. Critical t at 0.05/3 two-tailed, df = 9: 2.93. Same verdict for A vs B.

READ IT

The overall F was significant, but only A versus B separates. A significant F never means every pair differs.

TRAPS PEOPLE FALL INTO

- **Feeding percentages into χ^2.** Chi-square needs raw counts.
- **Rounding expected counts.** Keep E = 47.95 as 47.95.
- **Checking observed counts for the 5 rule.** The rule is about expected counts; an observed 0 is fine.
- **Reading a significant χ^2 as cause.** Survey tables show association only.
- **Running pairwise t-tests instead of ANOVA.** Each extra test adds false-alarm risk.
- **Writing Ha as "all means differ."** It is "at least one differs."
- **Pooling very unequal variances.** If the largest s is over twice the smallest, use Welch's ANOVA.
- **Fishing after a non-significant F.** Post-hoc tests follow a significant F, or planned contrasts set in advance.

Chapter 13 self-test · ten items

1. A claim says 50/30/20 for three categories, n = 150. What are the expected counts?

A. 50, 30, 20

B. 50, 50, 50

C. 100, 60, 40

D. 75, 45, 30

D. 75, 45, 30 E = n × p: 150(0.5) = 75, 150(0.3) = 45, 150(0.2) = 30.

2. O = 30 and E = 25 in one cell. What does it add to χ^2?

A. 1.00

B. 0.20

C. 5.00

D. 25.00

A. 1.00 $(30 - 25)^2 \div 25 = 25 \div 25 = 1.00$.

3. A goodness-of-fit test has 5 categories. What is df?

A. 5

B. 4

C. 6

D. 25

B. 4 df = k − 1 = 4.

4. In a two-way table, a cell has row total 80, column total 60 and grand total 200. What is E?

A. 24
B. 140
C. 40
D. 30

A. 24 (80 × 60) ÷ 200 = 4,800 ÷ 200 = 24.

5. A 3 × 4 table is tested for independence. What is df?

A. 12
B. 7
C. 11
D. 6

D. 6 (3 − 1)(4 − 1) = 2 × 3 = 6.

6. χ^2 = 12.0 from a 2 × 3 table with n = 400. What is Cramér's V?

A. 0.030
B. 0.122
C. 0.173
D. 0.300

C. 0.173 min(r − 1, c − 1) = 1, so V = √(12 ÷ 400) = √0.03 ≈ 0.173.

7. Why use ANOVA instead of three separate t-tests?

A. It always gives a larger F than a t-test
B. It does not need any conditions at all
C. It keeps the overall false-alarm rate at α
D. It shows directly which pair of means differs

C. It keeps the overall false-alarm rate at α Several t-tests inflate the family-wise error rate; one F test keeps it at α.

8. $SS_{between}$ = 90 with k = 4 groups and SS_{within} = 240 with N = 28. What is F?

A. 3.00
B. 0.38
C. 2.62
D. 30.00

A. 3.00 MSB = 90 ÷ 3 = 30; MSW = 240 ÷ 24 = 10; F = 3.00.

9. Group SDs are 4.0, 5.5 and 9.6. What does the equal-variance check say?

A. It passes, since all SDs are positive
B. It fails, since 9.6 is over twice 4.0
C. It passes, since the mean SD is 6.4
D. It cannot be checked without n

B. It fails, since 9.6 is over twice 4.0 9.6 ÷ 4.0 = 2.4, above the rule-of-thumb limit of 2. Consider Welch's ANOVA.

10. An ANOVA rejects H0. Tukey HSD = 6.0, and the means are 50, 54 and 61. Which pairs differ?

A. All three pairs differ
B. Only 50 vs 54 differs
C. None of the pairs differ
D. 50 vs 61 and 54 vs 61

D. 50 vs 61 and 54 vs 61 Gaps: 4, 11 and 7. Only gaps above 6.0 are significant.

CHAPTER 14

Correlation and Regression

Read it in full: **Chapter 14 in the textbook**. Practice it in the lab: **Read a Scatterplot** and **Fit the Least-Squares Line**.

What to know

A **scatterplot** shows two quantitative variables measured on the same individuals, one dot per individual. The **explanatory** variable x goes on the horizontal axis and the **response** y on the vertical. Describe it in order: **form** (linear, curved, clusters), **direction** (positive or negative), **strength** (how tightly points hug the form) and **outliers** (points off the joint pattern). Strength is not steepness: a shallow, tight band is strong.

The **correlation** $r = \Sigma(x - \bar{x})(y - \bar{y}) \div \sqrt{[\Sigma(x - \bar{x})^2 \times \Sigma(y - \bar{y})^2]}$, or the average product of z-scores, $\Sigma z_x z_y \div (n - 1)$. It lies between −1 and +1, has no units, is unchanged by rescaling and is symmetric in x and y. For the book's six students (hours 2 to 10, scores 72 to 94), $r = 126 \div \sqrt{(42 \times 466)} = 0.90$. Strength is distance from zero, so −0.87 is stronger than +0.60.

Four cautions: r measures **linear** association only (r near 0 can hide a strong curve); r is **not resistant** (Anscombe's quartet shares one r and one line across four very different pictures); a **restricted range** shrinks r; and **correlation is not causation**, since a lurking variable can produce any r. Only a randomized experiment licenses a causal claim.

The **least-squares line** $\hat{y} = a + bx$ minimizes $\Sigma(y - \hat{y})^2$, the sum of squared vertical distances. $b = \Sigma(x - \bar{x})(y - \bar{y}) \div \Sigma(x - \bar{x})^2 = r(s_y \div s_x)$ and $a = \bar{y} - b\bar{x}$, so the line always passes through $(\bar{x}, \bar{y})$ and the slope shares the sign of r. For the students, $b = 126 \div 42 = 3$ and $a = 78 - 18 = 60$: $\hat{y} = 60 + 3x$. Any other slope through (6, 78) gives a bigger SSE (98.5 for slopes 2.5 and 3.5 versus 88). Regression is not symmetric: predicting x from y needs its own line.

Interpret the **slope** as a predicted average rate: "for each additional hour, the model predicts about 3 more points, on average," not a causal promise. The **intercept** is the prediction at $x = 0$, meaningful only if $x = 0$ lies within the data; here hours ran from 2 to 10, so 60 just anchors the line.

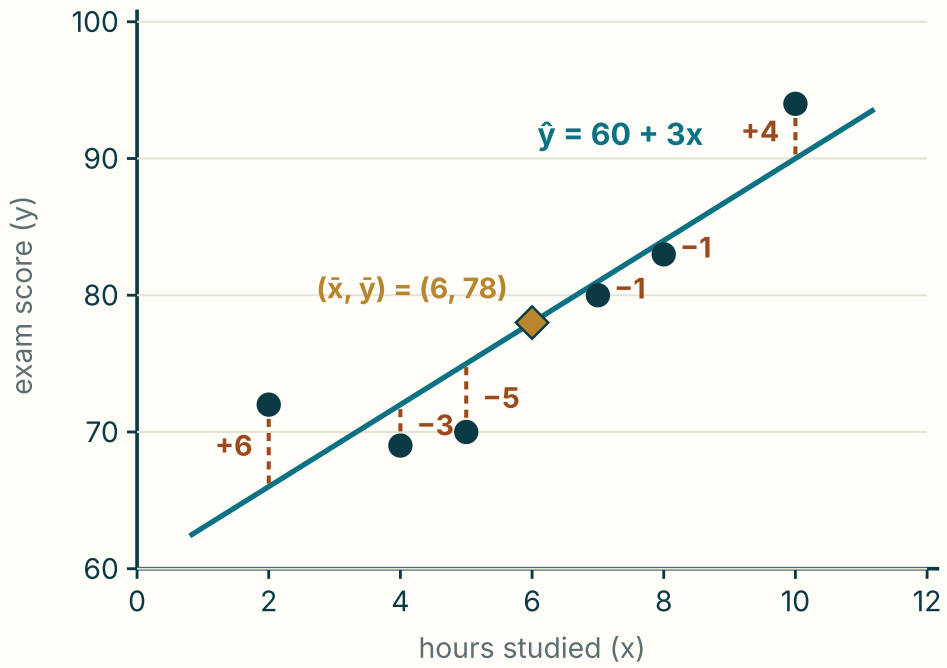
A **residual** is $y - \hat{y}$; least-squares residuals always sum to 0. $SST = \Sigma(y - \bar{y})^2 = 466$ and $SSE = \Sigma(y - \hat{y})^2 = 88$, so $\mathbf{r^2} = 1 - SSE \div SST = 0.81$: about 81% of the variation in scores is explained by the linear relationship with hours. It is not 81% of points on the line, not 81% accuracy and not 81% caused. r^2 hides direction, so report r too.

A **residual plot** (residuals against x or $\hat{y}$) should be a formless horizontal band. A U or arch means curvature; a fan means non-constant spread; a lone stray point is an outlier. **Leverage** is extremeness in x; a point is **influential** if removing it noticeably moves the slope, intercept or r. A high r^2 does not prove the model is right.

The line predicts the **average** y at a given x. The **standard error of the estimate** $s_e = \sqrt{[SSE \div (n - 2)]} = \sqrt{22} = 4.69$ is the typical miss (two df are spent on slope and intercept). **Interpolation** within the x range is supported; **extrapolation** is not: the line predicts an impossible 105 at 15 hours. Because $|r| < 1$, predictions regress toward $\bar{y}$. Inference on the slope assumes linearity, independence, constant variance and roughly normal residuals; a p-value is not the probability the slope is zero, and a 95% CI describes the procedure.

LOCK THESE IN

- Form, direction, strength, outliers; x is explanatory.
- r lies in [−1, 1], has no units and measures linear association only.
- $b = \Sigma(x - \bar{x})(y - \bar{y}) \div \Sigma(x - \bar{x})^2 = r\, s_y/s_x$; $a = \bar{y} - b\bar{x}$.
- The line passes through $(\bar{x}, \bar{y})$; residuals sum to 0.
- $r^2 = 1 - SSE/SST$ = share of variation in y explained by the line.
- $s_e = \sqrt{[SSE \div (n - 2)]}$.
- Interpolate, do not extrapolate. Refit to predict x from y.
- Correlation is not causation; strength is not steepness.



r = 0.90
r² = 0.81
SSE = 88
SST = 466
se = √(88 ÷ 4) = 4.69
residuals sum to 0

dashed: residual
y − ŷ

Least squares minimizes the squared vertical misses. The line ŷ = 60 + 3x passes through (6, 78). Its residuals +6, −3, −5, −1, −1 and +4 square and sum to SSE = 88, the smallest possible, giving $r^2 = 0.81$.

Formulas and when to use them

Quantity	Formula	When to use it
Correlation	$r = Sxy \div \sqrt{(Sxx \times Syy)}$	Strength and direction of a linear pattern
Correlation (z form)	$r = \Sigma z_x z_y \div (n - 1)$	Why r has no units
Slope	$b = Sxy \div Sxx = r(s_y \div s_x)$	Predicted change in y per unit x
Intercept	$a = \bar{y} - b\bar{x}$	Anchors the line
Prediction	$\hat{y} = a + bx$	Average y at x, inside the data range
Residual	$e = y - \hat{y}$	What the line missed
Sums of squares	$SST = \Sigma(y - \bar{y})^2$; $SSE = \Sigma(y - \hat{y})^2$	Variation before and after the line
r^2	$1 - SSE \div SST = r^2$	Share of variation explained
Std. error of estimate	$s_e = \sqrt{[SSE \div (n - 2)]}$	Typical miss

Reading a residual plot

What you see	What it means	What to do
Formless band around 0	A line fits	Report slope, r and r^2
U or arch	Curved form	Transform or fit a curve
Fan widening	Spread grows with x	Predictions weaker at the wide end
One point far off vertically	Outlier in y	Investigate; report with and without
One point far out in x	High leverage, maybe influential	Refit without it and compare

Worked examples

WORKED EXAMPLE 1 · COMPUTING R FROM A DEVIATION TABLE

Six weeks of ad spend x (hundreds of dollars) and sales y (thousands): x = 3, 5, 6, 8, 9, 11; y = 20, 25, 24, 31, 33, 37. Find r.

STEP 1 · MEANS

$\bar{x} = 42 \div 6 = 7.00$; $\bar{y} = 170 \div 6 = 28.33$.

STEP 2 · DEVIATIONS

$x - \bar{x}$: −4.00, −2.00, −1.00, 1.00, 2.00, 4.00. $y - \bar{y}$: −8.33, −3.33, −4.33, 2.67, 4.67, 8.67. Both sum to 0.

STEP 3 · SUMS

$S_{xy} = \Sigma(x - \bar{x})(y - \bar{y}) = 91.00$; $S_{xx} = 42.00$; $S_{yy} = 203.33$.

STEP 4 · R

$r = 91.00 \div \sqrt{(42.00 \times 203.33)} = 91.00 \div 92.41 =$ **0.985**.

STEP 5 · READ IT

Strong and positive, and only after the scatterplot shows a straight band. These are observational weeks, so r says nothing about whether ads caused the sales.

WORKED EXAMPLE 2 · FITTING AND INTERPRETING THE LINE

Fit the least-squares line for the ad data and interpret it.

STEP 1 · SLOPE

$b = S_{xy} \div S_{xx} = 91.00 \div 42.00 =$ **2.1667**. Check: $r(s_y \div s_x) = 0.9847 \times (6.3770 \div 2.8983) = 2.1667$.

STEP 2 · INTERCEPT

$a = 28.33 - 2.1667(7.00) =$ **13.167**.

STEP 3 · THE LINE

$\hat{y} = 13.17 + 2.167x$. At $\bar{x} = 7.00$, $\hat{y} = 28.33 = \bar{y}$.

STEP 4 · SLOPE IN WORDS

Each extra $100 of ad spend goes with about 2.17 thousand dollars more in predicted sales, on average.

STEP 5 · INTERCEPT IN WORDS

x = 0 is outside the observed 3 to 11, so 13.17 only anchors the line; it is not a forecast for zero spending.

WORKED EXAMPLE 3 · RESIDUALS, R^2 AND THE STANDARD ERROR OF THE ESTIMATE

Find the residuals, SSE, r^2 and s_e for the ad data.

STEP 1 · PREDICTIONS

19.67, 24.00, 26.17, 30.50, 32.67, 37.00.

STEP 2 · RESIDUALS

0.33, 1.00, −2.17, 0.50, 0.33, 0.00. They sum to 0.00.

STEP 3 · SSE

Σ residual² = **6.167**; SST = Syy = 203.33.

STEP 4 · R^2

1 − 6.167 ÷ 203.33 = **0.970**, matching $r^2 = 0.985^2 = 0.970$. About 97% of the week-to-week variation in sales is explained by the linear relationship with ad spend.

STEP 5 · S_E

√(6.167 ÷ 4) = **1.24** thousand dollars: a typical week sits about that far from the line.

WORKED EXAMPLE 4 · PREDICTING, AND WHERE NOT TO

Use the ad line to predict sales at x = 7 and x = 20, and explain the difference. Then predict hours studied from a score of 90 in the book's student data.

STEP 1 · INTERPOLATE

$\hat{y} = 13.17 + 2.167(7)$ = **28.33**, inside the observed 3 to 11; give or take about $2s_e = 2.5$.

STEP 2 · EXTRAPOLATE

$\hat{y} = 13.17 + 2.167(20) = 56.50$. x = 20 is far beyond the data; the line cannot know whether sales level off, so this number is not supported.

STEP 3 · REVERSE PREDICTION

Inverting $\hat{y} = 60 + 3x$ gives (90 − 60) ÷ 3 = 10 hours, but that is the wrong line. Refit with hours as the response: slope = 126 ÷ 466 = **0.270**, intercept = 6 − 0.270(78) = −15.09, so predicted hours = 9.24.

WHY THEY DIFFER

Each regression pulls its predictions toward the mean of its own response. With $|r| < 1$, the two lines are not inverses.

TRAPS PEOPLE FALL INTO

- **Computing r before looking.** r is blind to curves, clusters and single influential points.
- **Treating strength as steepness.** A shallow tight band is strong; a steep fuzzy one is weak.
- **Calling −0.87 weaker than +0.60.** Strength is distance from zero.
- **Reading the slope as cause.** Say "predicted, on average" for observational data.
- **Interpreting an intercept outside the data.** If x = 0 was never observed, the intercept only anchors the line.
- **Saying r^2 is the share of points on the line.** It is the share of variation in y explained.
- **Extrapolating.** The line predicted an impossible 105 at 15 hours.

- **Inverting the line to predict x.** Refit with the roles reversed.

Chapter 14 self-test · ten items

1. Which variable belongs on the horizontal axis?

A. The response variable
B. The one with larger values
C. The explanatory variable
D. Either, it never matters

C. The explanatory variable x is the explanatory variable; the regression line depends on that choice.

2. Which r shows the strongest linear relationship?

A. +0.85
B. +0.10
C. −0.40
D. −0.92

D. −0.92 Strength is distance from zero, regardless of sign.

3. Sxy = 60, Sxx = 20 and Syy = 250. What is r?

A. 0.849
B. 3.000
C. 0.240
D. 0.012

A. 0.849 r = 60 ÷ √(20 × 250) = 60 ÷ √5,000 ≈ 0.849.

4. Using the same sums (Sxy = 60, Sxx = 20) with $\bar{x} = 5$ and $\bar{y} = 40$, what is the line?

A. $\hat{y} = 40 + 3x$
B. $\hat{y} = 25 + 3x$
C. $\hat{y} = 5 + 3x$
D. $\hat{y} = 25 + 0.33x$

B. $\hat{y} = 25 + 3x$ b = 60 ÷ 20 = 3; a = 40 − 3(5) = 25.

5. r = 0.6, $s_y = 10$, $s_x = 4$. What is the slope?

A. 1.5
B. 0.24
C. 6.0
D. 2.4

A. 1.5 $b = r(s_y \div s_x) = 0.6 \times 2.5 = 1.5$.

6. $\hat{y} = 12 + 4x$ and a point is (3, 21). What is its residual?

A. 3
B. −3
C. 24
D. 9

B. −3 $\hat{y} = 12 + 12 = 24$; residual = 21 − 24 = −3.

7. SST = 500 and SSE = 125. What is r^2?

A. 0.75
B. 0.25
C. 4.00
D. 0.50

A. 0.75 1 − 125 ÷ 500 = 0.75.

8. $r^2 = 0.64$ for a regression. Which statement is correct?

A. 64% of the points lie exactly on the line
B. Predictions from the line are 64% accurate
C. x causes 64% of the changes in y
D. 64% of the variation in y is explained by the line

D. 64% of the variation in y is explained by the line r^2 is a statement about variation, nothing else. Here r = ±0.8 depending on direction.

9. A residual plot shows a clear U shape. What does it mean?

A. The straight line fits the data well
B. The true form is curved, so a line is a poor model
C. The spread of the residuals is constant
D. There are no outliers in the data

B. The true form is curved, so a line is a poor model Curvature left after removing the line means the linear model misses the form.

10. Data cover x from 20 to 60. Which prediction is extrapolation?

A. x = 40

B. x = 25

C. x = 58

D. x = 85

D. x = 85 Only x = 85 lies outside the observed range.

ONE PAGE

Cram sheet

Everything from chapters 13 and 14 in one place. Cover the right column and see if you can rebuild each row.

Formulas

Name	Formula	Note
Chi-square	$\Sigma(O - E)^2/E$	Right-tailed
Expected counts	np; row × column ÷ total	All ≥ 5; never round
df	$k - 1$; $(r - 1)(c - 1)$	GOF; independence
Cramér's V	$\sqrt{[\chi^2/(n \times \min(r - 1, c - 1))]}$	Size of association
ANOVA	$F = MSB/MSW$; df $k - 1$, $N - k$	$SS_T = SS_B + SS_W$
Effect size	$\eta^2 = SS_B/SS_T$	Share explained
Tukey	$HSD = q\sqrt{(MSW/n)}$	Bonferroni: α/m
Correlation	$r = Sxy/\sqrt{(Sxx\ Syy)}$	$[-1, 1]$, unit-free
Line	$b = Sxy/Sxx = r\ s_y/s_x$; $a = \bar{y} - b\bar{x}$	Through $(\bar{x}, \bar{y})$
Fit	$r^2 = 1 - SSE/SST$; $s_e = \sqrt{[SSE/(n - 2)]}$	Residuals sum to 0

Rules that are not formulas

Topic	The rule
Chi-square data	Counts only; paired yes/no uses McNemar.
Small counts	Combine categories or use Fisher's exact test.
ANOVA Ha	At least one mean differs.
ANOVA conditions	Independence, rough normality, largest s < 2 × smallest.
After F	Post-hoc only after a significant F (or planned contrasts).
Scatterplot	Form, direction, strength, outliers.
r limits	Linear only, not resistant, shrinks with restricted range, not causal.
Residual plot	Want no pattern; U = curve; fan = changing spread.
Prediction	Interpolate only; refit to predict x from y.

Dr. Garg Learning · CC BY-SA 4.0