UNIVERSITY OF LEEDS

SCHOOL OF MATHEMATICS

STATISTICAL QUESTION BANK

University-Level Dataset | Categorised A–F |
50 Curated Problems | No Solutions Included

Statistics Questions Dataset - Marker Guidance (Questions Only)

Overview

This document contains 50 statistics questions organized using an A–F categorization system. This is the questions-only version. No solutions are provided.

★ For worked-out answers and marking guidance, please refer to the companion document: Solution Bank.

.Dataset Structure

Indexing System

- **Format**: Letter-Level-Number (e.g., A1.234, B3.789, F3.901)
- **Letters**: Represent topic categories (A through F)
- **Numbers**: Simulated indices from a larger 2000-question database
- **Sample Size**: 50 questions strategically selected to represent the full spectrum

Category Framework

Code	Category	Description	Question Count
A	Probability & Bayes	Binomial, Poisson, Normal distributions, conditional probability, Markov chains	19 (38%)
B	Exploratory Data Analysis	Data visualization, descriptive statistics, gender/demographic analysis	1 (2%)
C	Confidence Intervals	Population means, proportions, sample size calculations	6 (12%)
D	Parameter Estimation	Sampling distributions, linear regression, Central Limit Theorem	4 (8%)
E	Hypothesis Testing	ANOVA, t-tests, z-tests, proportions, non-parametric methods	12 (24%)
F	Bayesian Inference	Bayes' theorem, posterior analysis, logistic regression	8 (16%)

```
 "questions": [
  {
   "index": "A1.234",
   "category": "A",
   "difficulty_level": 1,
   "difficulty_name": "Basic",
   "topic": "Binomial Distribution",
   "problem": "A ball is drawn from an urn containing three white and three black balls. After
the ball is drawn, it is then replaced and another ball is drawn. This goes on indefinitely. What is
the probability that of the first four balls drawn, exactly two are white?"
  },
  {
   "index": "A1.567",
   "category": "A",
   "difficulty_level": 1,
   "difficulty_name": "Basic",
   "topic": "Probability",
   "problem": "A game show has a prize with a 5% chance of being won. If a contestant plays 7
games, what is the probability of winning the prize at least once?"
  },
  {
   "index": "A2.123",
   "category": "A",
   "difficulty_level": 2,
   "difficulty_name": "Intermediate",
   "topic": "Conditional Probability",
   "problem": "Among a group of 200 students, 137 students are enrolled in a mathematics
class, 50 in a history class, and 124 in a music class. Additionally, 33 students are enrolled in
both mathematics and history, 29 in both history and music, and 92 in both mathematics and
music. Finally, 18 students are enrolled in all three classes. What is the probability that a
randomly selected student is enrolled in at least one of the three classes?"
  },
  {
   "index": "A2.445",
   "category": "A",
   "difficulty_level": 2,
   "difficulty_name": "Intermediate",
   "topic": "Binomial Distribution",
   "problem": "Approximately 15% of the US population smokes cigarettes. A local government
believed their community had a lower smoker rate and commissioned a survey of 400 randomly
selected individuals. The survey found that only 42 of the 400 participants smoke cigarettes. If
```

```
the true proportion of smokers in the community was really 15%, what is the probability of
observing 42 or fewer smokers in a sample of 400 people?"
  },
  {
    "index": "A2.678",
    "category": "A",
    "difficulty_level": 2,
    "difficulty_name": "Intermediate",
    "topic": "Probability",
    "problem": "A factory produces 1000 light bulbs, of which 40 are defective. If 5 bulbs are
randomly selected without replacement, what is the probability that exactly 2 of them are
defective?"
  },
  {
    "index": "A2.789",
    "category": "A",
    "difficulty_level": 2,
    "difficulty_name": "Intermediate",
    "topic": "Binomial Distribution",
    "problem": "At Selitall Supermarket, 60% of customers pay by credit card. Find the
probability that in a randomly selected sample of ten customers, (a) exactly two pay by credit
card, (b) more than seven pay by credit card."
  },
  {
    "index": "A2.345",
    "category": "A",
    "difficulty_level": 2,
    "difficulty_name": "Intermediate",
    "topic": "Poisson Distribution",
    "problem": "Two identical racing cars are being tested on a circuit. For each car, the number
of mechanical breakdowns can be modelled by a Poisson distribution with a mean of one
breakdown in 100 laps. If a car breaks down, it is attended and continues on the circuit. The first
car is tested for 20 laps and the second car for 40 laps. Find the probability that the service
team is called out to attend to breakdowns: (a) once, (b) more than twice."
  },
  {
    "index": "A2.456",
    "category": "A",
    "difficulty_level": 2,
    "difficulty_name": "Intermediate",
    "topic": "Normal Distribution",
    "problem": "Lengths of metal strips produced by a machine are normally distributed with a
mean length of 150 cm and a standard deviation of 10 cm. Find the probability that the length of
a randomly selected strip is: (a) shorter than 165 cm, (b) within 5 cm of the mean."
```

```
  },
  {
    "index": "A2.567",
    "category": "A",
    "difficulty_level": 2,
    "difficulty_name": "Intermediate",
    "topic": "Markov Chains",
    "problem": "A cat and a mouse move independently back and forth between two rooms. At each time step, the cat moves from the current room to the other room with probability 0.8. Starting from room 1, the mouse moves to room 2 with probability 0.3 (and remains otherwise). Starting from room 2, the mouse moves to room 1 with probability 0.6 (and remains otherwise). (a) Find the stationary distributions of the cat chain and of the mouse chain."
  },
  {
    "index": "A2.678",
    "category": "A",
    "difficulty_level": 2,
    "difficulty_name": "Intermediate",
    "topic": "Discrete Probability",
    "problem": "Five people have just won a $100 prize and are deciding how to divide the $100 up between them. Assume that whole dollars are used, not cents. Also, for example, giving $50 to the first person and $10 to the second is different from vice versa. (a) How many ways are there to divide up the $100, such that each person gets at least $10?"
  },
  {
    "index": "A2.789",
    "category": "A",
    "difficulty_level": 2,
    "difficulty_name": "Intermediate",
    "topic": "Conditional Probability",
    "problem": "A game show presents contestants with four doors: behind one is a car worth $1000, behind another is a forfeit costing -$1000, and behind the remaining two doors is nothing. The contestant picks one door. The host then opens one of the remaining doors to show it is empty. The contestant may stick with their choice or switch to one of the other two unopened doors. What is the optimal strategy if the contestant wants to maximise expected wealth and is risk-averse?"
  },
  {
    "index": "A2.890",
    "category": "A",
    "difficulty_level": 2,
    "difficulty_name": "Intermediate",
    "topic": "Conditional Probability – Boy or Girl Paradox",
    "problem": "Mr Bayes has two children. The older child is a girl. What is the probability that both children are girls?"
  },
  {
    "index": "A2.901",
    "category": "A",
    "difficulty_level": 2,
    "difficulty_name": "Intermediate",
    "topic": "Conditional Probability – Boy or Girl Paradox",
    "problem": "Mr Laplace has two children. At least one of the children is a girl. What is the probability that both children are girls?"
  },
  {
    "index": "A3.123",
    "category": "A",
    "difficulty_level": 3,
    "difficulty_name": "Advanced",
    "topic": "Probability",
    "problem": "We roll a fair four-sided die. If the result is 1 or 2, we roll once more but otherwise, we stop. What is the probability that the sum total of our rolls is at least 4?"
  },
  {
    "index": "A3.234",
    "category": "A",
    "difficulty_level": 3,
    "difficulty_name": "Advanced",
    "topic": "Probability",
    "problem": "A store has 20% off sales on items. If a customer buys 8 items, what is the probability that at least 3 of the items are on sale?"
  },
  {
    "index": "A3.345",
    "category": "A",
    "difficulty_level": 3,
    "difficulty_name": "Advanced",
    "topic": "Probability",
    "problem": "In a population of 200 people, 30 have a certain disease. If 10 people are randomly selected, what is the probability that exactly 4 of them have the disease?"
  },
  {
    "index": "A3.456",
    "category": "A",
    "difficulty_level": 3,
    "difficulty_name": "Advanced",
    "topic": "Bias and Logistic Thinking",
    "problem": "Let's consider a sex-imbalanced company that consists of 20% women and 80% men. Suppose that the company is very large, consisting of perhaps 20,000 employees. Suppose when someone goes up for promotion at this company, 5 of their colleagues are randomly chosen to provide feedback on their work. Now let's imagine that 10% of the people in the company are prejudiced against the other sex. That is, 10% of men are prejudiced against women, and similarly, 10% of women are prejudiced against men. Who is discriminated against more at the company, men or women?"
  },
  {
    "index": "A3.567",
    "category": "A",
    "difficulty_level": 3,
    "difficulty_name": "Advanced",
    "topic": "Probability Distributions",
    "problem": "Heights of 10 year olds, regardless of gender, closely follow a normal distribution with mean 55 inches and standard deviation 6 inches.",
    "questions": [
      "(a) What fraction of 10 year olds are taller than 76 inches?",
      "(b) If there are 2,000 10 year olds entering Six Flags Magic Mountain in a single day, then compute the expected number of 10 year olds who are at least 76 inches tall. (Assume independence.)",
      "(c) Using the binomial distribution, compute the probability that 0 of the 2,000 10 year olds will be at least 76 inches tall.",
      "(d) The number of 10 year olds who enter Six Flags Magic Mountain and are at least 76 inches tall in a given day follows a Poisson distribution with mean equal to the value found in part (b). Use the Poisson distribution to identify the probability that no 10 year old will enter the park who is 76 inches or taller."
    ]
  },
  {
    "index": "A3.678",
    "category": "A",
    "difficulty_level": 3,
    "difficulty_name": "Advanced",
    "topic": "Poisson Distribution with Uncertainty",
    "problem": "The opponents of soccer team A are of two types: either they are a class 1 or a class 2 team. The number of goals team A scores against a class i opponent is a Poisson random variable with mean λᵢ, where λ₁ = 2, λ₂ = 3. This weekend, the team has two games against teams they are not very familiar with. Assuming that the first team they play is a class 1 team with probability 0.6 and the second is, independently of the class of the first team, a class 1 team with probability 0.3, determine: (a) the expected number of goals team A will score this weekend; (b) the probability that team A will score a total of five goals."
  },
  {
    "index": "B3.789",
    "category": "B",
    "difficulty_level": 3,
    "difficulty_name": "Advanced",
    "topic": "Gender and Degree Distributions",
    "problem": "Using the provided counts (in thousands) of earned degrees in the U.S. by gender and degree type, perform an exploratory data analysis to understand trends and imbalances. Specifically, analyze: (1) overall gender distribution, (2) the distribution of degrees within each gender, (3) the conditional probability of selecting a male given a Master's degree, (4) gender gap index for each degree type, and (5) the most gender-skewed degree type.",
    "data_summary": {
      "Bachelor's": {"Female": 616, "Male": 529},
      "Master's": {"Female": 194, "Male": 171},
      "Professional": {"Female": 30, "Male": 44},
      "Doctorate": {"Female": 16, "Male": 26}
    }
  },
  {
    "index": "C1.234",
    "category": "C",
    "difficulty_level": 1,
    "difficulty_name": "Basic",
    "topic": "Confidence Interval",
    "problem": "A teacher records the test scores of 10 students with an average score of 78 and a standard deviation of 5. Calculate the 95% confidence interval for the mean test score."
  },
  {
    "index": "C1.456",
    "category": "C",
    "difficulty_level": 1,
    "difficulty_name": "Basic",
    "topic": "Confidence Interval",
    "problem": "A sample of 18 students has an average test score of 72 with a standard deviation of 8. Calculate the 95% confidence interval for the average test score."
  },
  {
    "index": "C1.789",
    "category": "C",
    "difficulty_level": 1,
    "difficulty_name": "Basic",
    "topic": "Population Proportion",
    "problem": "A market research firm is hired to estimate the proportion of adults in a large city who own smartphones. A random sample of 500 adult residents is surveyed, and 421 say they
```

```
own smartphones. Using a 95% confidence level, compute a confidence interval for the true proportion of adult residents who own smartphones."
  },
  {
    "index": "C2.123",
    "category": "C",
    "difficulty_level": 2,
    "difficulty_name": "Intermediate",
    "topic": "Confidence Interval for a Mean",
    "problem": "The General Social Survey asked the question: 'For how many days during the past 30 days was your mental health, which includes stress, depression, and problems with emotions, not good?' Based on responses from 1,151 US residents, the survey reported a 95% confidence interval of 3.40 to 4.24 days in 2010.",
    "questions": [
      "(a) Interpret this interval in context of the data.",
      "(b) What does '95% confident' mean? Explain in the context of the application.",
      "(c) Suppose the researchers think a 99% confidence level would be more appropriate for this interval. Will this new interval be smaller or wider than the 95% confidence interval?",
      "(d) If a new survey were to be done with 500 Americans, do you think the standard error of the estimate would be larger, smaller, or about the same?"
    ]
  },
  {
    "index": "C2.345",
    "category": "C",
    "difficulty_level": 2,
    "difficulty_name": "Intermediate",
    "topic": "Confidence Interval",
    "problem": "A manager is about to oversee the mass production of a new tire model in her factory, and she would like to estimate what proportion of these tires will be rejected through quality control. The quality control team has monitored the last three tire models produced by the factory, failing 1.7% of tires in the first model, 6.2% of the second model, and 1.3% of the third model. The manager would like to examine enough tires to estimate the failure rate of the new tire model to within about 1% with a 90% confidence level. Perform the sample size computation for each past model separately."
  },
  {
    "index": "C3.567",
    "category": "C",
    "difficulty_level": 3,
    "difficulty_name": "Advanced",
    "topic": "Confidence Intervals",
    "problem": "You are given a dataset assumed to be i.i.d. from a normal distribution: X₁, ..., X□ ~ N(μ, σ²). Let τ be the 95th percentile of the distribution, i.e., P(X < τ) = 0.95. Suppose the
```

observed data are: 3.23, -2.50, 1.88, -0.68, 4.43, 0.17, 1.03, -0.07, -0.01, 0.76, 1.76, 3.18, 0.33, -0.31, 0.30, -0.61, 1.52, 5.43, 1.54, 2.28, 0.42, 2.33, -1.03, 4.00, 0.39",
"tasks": [
"1. Find the MLE of τ (the 95th percentile).",
"2. Find the standard error of the MLE using the delta method.",
"3. Estimate the standard error using the parametric bootstrap."
]
},
{
"index": "D1.234",
"category": "D",
"difficulty_level": 1,
"difficulty_name": "Basic",
"topic": "Sampling Distribution and Central Limit Theorem",
"problem": "Suppose that the height of men has mean 68 inches and standard deviation 2.6 inches. We draw a random sample of 100 men. Find (approximately) the probability that the average height of men in our sample will be at least 68 inches."
},
{
"index": "D2.345",
"category": "D",
"difficulty_level": 2,
"difficulty_name": "Intermediate",
"topic": "Linear Regression",
"problem": "A survey of 55 Duke University students asked about their GPA, number of hours they study at night (studyweek), number of nights they go out (outnight), number of hours they sleep at night (sleepnight), and their gender. The regression summary is given with coefficients for each predictor. Note: gender is coded as 1 for male.",
"questions": [
"(a) Calculate a 95% confidence interval for the coefficient of gender in the model, and interpret it in the context of the data.",
"(b) Would you expect a 95% confidence interval for the slope of the remaining variables to include 0? Explain."
]
},
{
"index": "D2.456",
"category": "D",
"difficulty_level": 2,
"difficulty_name": "Intermediate",
"topic": "Simple Linear Regression",
"problem": "A real estate analyst is modeling house prices based on square footage. She fits a simple linear regression model: $Price_i = \beta_0 + \beta_1 * SqFt_i + \varepsilon_i$. Given a dataset with 50 houses,

the following statistics are known: $\sum X_i = 75000$, $\sum Y_i = 5000000$, $\sum X_iY_i = 80000000$, $\sum X_i^2 = 120000000$, $\sum Y_i^2 = 600000000000$.",
    "questions": [
      "(a) Estimate the regression coefficients β_0 and β_1.",
      "(b) Interpret the meaning of these coefficients in the context of the problem."
    ]
  },
  {
    "index": "D2.567",
    "category": "D",
    "difficulty_level": 2,
    "difficulty_name": "Intermediate",
    "topic": "Linear Regression Analysis and Crystal Growth",
    "problem": "Crystalline forms of certain chemical compounds are used in various electronic devices. It is often more desirable to have large crystals rather than small ones. In a laboratory study, 14 crystals of the same initial size were allowed to grow for certain periods of time. The following data gives the weight y of the crystal (in grams) and the period x of time (in hours) which was used for each crystal: Time: 2,4,6,8,10,12,14,16,18,20,22,24,26,28; Weight: 0.08,1.12,4.43,4.98,4.92,7.18,5.57,8.4,8.81,10.81,11.16,10.12,13.12,15.04.",
    "tasks": [
      "a. Construct a scatterplot of the 'y data' versus the 'x data'.",
      "b. Find the sample mean(s) of the weight(y) and the period (x) of time.",
      "c. Compute the least-squares estimates of β_0 and β_1.",
      "d. Find and draw the Least-Square regression line and use it to estimate the mean weight in grams for a period of x = 5 hours.",
      "e. Does the line pass through the data points?",
      "f. Determine the coefficient of determination for crystalline forms."
    ]
  },
  {
    "index": "E1.123",
    "category": "E",
    "difficulty_level": 1,
    "difficulty_name": "Basic",
    "topic": "Hypothesis Testing",
    "problem": "For each of the following assertions, state whether it is a legitimate statistical hypothesis and why: a. H_0: $\sigma = 100$ b. H_0: $\sigma \geq 45$ c. H_0: $\sigma \leq 20$ d. H_0: $\sigma_1 / \sigma_2 = 1$ e. H_0: $X^2 \leq 55$ f. H_0: $\lambda \neq 0.01$, where λ is the parameter of an exponential distribution used to model component lifetime."
  },
  {
    "index": "E1.234",
    "category": "E",
    "difficulty_level": 1,

```
      "difficulty_name": "Basic",
      "topic": "ANOVA",
      "problem": "A fitness trainer tests the effectiveness of three different workout programs (Program A, Program B, Program C) on weight loss. The weight loss (in pounds) for participants in each program is as follows: Program A = [3, 4, 5], Program B = [2, 3, 4], Program C = [5, 6, 7]. Perform a one-way ANOVA test to determine if there is a significant difference in weight loss among the programs."
    },
    {
      "index": "E2.345",
      "category": "E",
      "difficulty_level": 2,
      "difficulty_name": "Intermediate",
      "topic": "ANOVA",
      "problem": "A company tests whether different leadership styles (Style A, Style B, Style C) have different effects on employee productivity. The productivity scores for each style are as follows: Style A = [85, 88, 90], Style B = [78, 80, 82], Style C = [88, 90, 92]. Perform a one-way ANOVA test."
    },
    {
      "index": "E2.456",
      "category": "E",
      "difficulty_level": 2,
      "difficulty_name": "Intermediate",
      "topic": "ANOVA",
      "problem": "A nutritionist tests the effects of three different meal plans (Plan A, Plan B, Plan C) on weight loss over a 6-week period. The weight loss (in pounds) for each plan is as follows: Plan A = [6, 7, 8], Plan B = [5, 6, 7], Plan C = [8, 9, 10]. Conduct a one-way ANOVA test."
    },
    {
      "index": "E2.567",
      "category": "E",
      "difficulty_level": 2,
      "difficulty_name": "Intermediate",
      "topic": "Hypothesis Testing (Proportions)",
      "problem": "A newspaper reported that 75% of students regularly cycle to college. A college dean believes that figure to be different at his college. He asks a sample of 160 students, and 109 say they do cycle. Test the dean's belief at the 5% level of significance."
    },
    {
      "index": "E2.678",
      "category": "E",
      "difficulty_level": 2,
      "difficulty_name": "Intermediate",
      "topic": "Hypothesis Testing - Binomial to Normal Approximation",
      "problem": "In a test, the questions are all multiple choice with five possible options.",
      "questions": [
        {
          "part": "a",
          "description": "In a test with twelve questions, one student gets four questions correct. Test, at the 10% significance level, the null hypothesis that the student is guessing the answers."
        },
        {
          "part": "b",
          "description": "In a further test there are 120 questions. The same student took the test and got 32 correct. Test, at the 10% significance level, whether there is evidence to show the student is guessing the answers."
        }
      ]
    },
    {
      "index": "E2.789",
      "category": "E",
      "difficulty_level": 2,
      "difficulty_name": "Intermediate",
      "topic": "Hypothesis testing",
      "problem": "Georgianna claims that in a small city renowned for its music school, the average child takes less than 5 years of piano lessons. A random sample of 24 children from the city shows a mean of 4.6 years of piano lessons with a standard deviation of 2.2 years. Evaluate Georgianna's claim using a hypothesis test at a 0.05 significance level."
    },
    {
      "index": "E2.890",
      "category": "E",
      "difficulty_level": 2,
      "difficulty_name": "Intermediate",
      "topic": "One-Sample Z-Test (Known σ)",
      "problem": "Jeffrey, an eight-year-old, had an established mean time of 16.43 seconds for swimming the 25-yard freestyle, with a population standard deviation of 0.8 seconds. His father Frank bought him expensive goggles, suspecting they might help him swim faster. In 15 timed swims using the goggles, Jeffrey's sample mean time was 16 seconds. Test whether the goggles helped Jeffrey swim faster, using a significance level of α = 0.05. Assume the swim times are normally distributed."
    },
    {
      "index": "E3.234",
      "category": "E",
      "difficulty_level": 3,
      "difficulty_name": "Advanced",
      "topic": "One Sample t-test",
      "problem": "Use the data [0.86, 1.53, 1.57, 1.81, 0.99, 1.09, 1.29, 1.78, 1.29, 1.58], comprising a sample of n = 10 lactic acid measurements in cheese. The lactic acid measurements are a random sample from a normal distribution with unknown mean μ and unknown variance σ².",
      "hypotheses": {
        "null": "H₀: μ ≤ 1.2",
        "alternative": "H₁: μ > 1.2"
      },
      "tasks": [
        "a. Perform the level α₀ = 0.05 test of these hypotheses.",
        "b. Compute the p-value."
      ]
    },
    {
      "index": "E3.345",
      "category": "E",
      "difficulty_level": 3,
      "difficulty_name": "Advanced",
      "topic": "Hypothesis Testing and Nonparametric Inference",
      "problem": "In 1861, 10 essays appeared in the New Orleans Daily Crescent signed 'Quintus Curtius Snodgrass.' Some suspected they were actually written by Mark Twain. To investigate, we consider the proportion of three-letter words in the texts. From 8 Twain essays: [0.225, 0.262, 0.217, 0.240, 0.230, 0.229, 0.235, 0.217]. From 10 Snodgrass essays: [0.209, 0.205, 0.196, 0.210, 0.202, 0.207, 0.224, 0.223, 0.220, 0.201].",
      "questions": [
        "(a) Perform a Wald test for equality of the means. Use the nonparametric plug-in estimator (difference of the sample means). Report the p-value and a 95% confidence interval for the difference of means. What do you conclude?",
        "(b) Use a permutation test to avoid the large-sample approximation. What is your conclusion?"
      ]
    },
    {
      "index": "E3.456",
      "category": "E",
      "difficulty_level": 3,
      "difficulty_name": "Advanced",
      "topic": "Hypothesis Testing – Binomial Distribution",
      "problem": "A paper company finds that, over time, defects occur at a rate of 1 in every 250 sheets of paper produced (p = 0.004). After switching to recycled paper, the company examines a sample of 300 sheets and finds 5 defective sheets. An employee conducts a hypothesis test
```

at the 5% level of significance to assess whether the switch has affected the defect rate. The employee assumes X ~ B(300, p), sets H_0: p = 0.004 and H_1: p ≠ 0.004, and calculates P(X = 5 | p = 0.004) = 0.9985. Based on this, they conclude: 'Since 0.9985 > 0.05, accept H_0. There is no evidence that the defect rate has changed.' Explain the mistakes and determine the correct conclusion."
    },
    {
      "index": "E3.1567",
      "category": "E",
      "difficulty_level": 3,
      "difficulty_name": "Advanced",
      "topic": "One-Way ANOVA – Hospital ICU Hours",
      "problem": "The Eastside Health Authority collects data on hours spent in intensive care by patients with suspected coronary heart attacks across five hospitals (A–E). The data are as follows: Hospital A: 30, 25, 12, 23, 16; Hospital B: 42, 57, 47, 30; Hospital C: 65, 46, 55, 27; Hospital D: 67, 58, 81; Hospital E: 70, 63, 80. Use a one-factor analysis of variance to test, at the 1% level of significance, whether there is a difference in mean ICU hours between hospitals."
    },
    {
      "index": "F1.234",
      "category": "F",
      "difficulty_level": 1,
      "difficulty_name": "Basic",
      "topic": "Bayes' Theorem",
      "problem": "A spam filter is designed by looking at commonly occurring phrases in spam. Suppose that 80% of email is spam. In 10% of the spam emails, the phrase 'free money' is used, whereas this phrase is only used in 1% of non-spam emails. A new email has just arrived, which does mention 'free money'. What is the probability that it is spam?"
    },
    {
      "index": "F2.345",
      "category": "F",
      "difficulty_level": 2,
      "difficulty_name": "Intermediate",
      "topic": "Bayesian Statistics",
      "problem": "Given a Beta(5, 3) prior and observing 25 successes and 10 failures, find the posterior distribution parameters and compute the posterior variance."
    },
    {
      "index": "F3.456",
      "category": "F",
      "difficulty_level": 3,
      "difficulty_name": "Advanced",

```
      "topic": "Bayes' Theorem – Inverted Probability",
      "problem": "In Canada, about 0.35% of women over 40 will develop breast cancer in any given year. A common screening test for cancer is the mammogram, but it is not perfect. In 11% of patients with breast cancer, the test gives a false negative result. In 7% of patients without breast cancer, the test gives a false positive result. If we test a random woman over 40 for breast cancer using a mammogram and the test comes back positive, what is the probability that the patient actually has breast cancer?"
    },
    {
      "index": "F3.567",
      "category": "F",
      "difficulty_level": 3,
      "difficulty_name": "Advanced",
      "topic": "Logistic Regression",
      "problem": "A logistic regression model was fit to estimate the probability of receiving a callback based on whether a resume listed any honors. The model equation is: log_e(p_i / (1 - p_i)) = -2.4998 + 0.8668 × honors.",
      "questions": [
        "(a) If a resume does not have honors listed, what is the probability of getting a callback?",
        "(b) What is the probability if the resume does list honors?"
      ]
    },
    {
      "index": "F3.678",
      "category": "F",
      "difficulty_level": 3,
      "difficulty_name": "Advanced",
      "topic": "Logistic Regression",
      "problem": "A logistic regression model is fitted using a binary predictor: whether an applicant had any type of honors listed on their resume (e.g., 'employee of the month'). The model is loge(p / (1 - p)) = -2.4998 + 0.8668 × honors.",
      "questions": [
        "(a) What is the callback probability if a resume has no honors?",
        "(b) What is the probability if the resume includes honors?"
      ]
    },
    {
      "index": "F3.789",
      "category": "F",
      "difficulty_level": 3,
      "difficulty_name": "Advanced",
      "topic": "Bayesian Statistics",
      "problem": "Assume a Gamma(4, 5) prior and observe 15 successes and 10 failures. Compute the posterior distribution parameters and posterior mean."
    },
    {
      "index": "F3.890",
      "category": "F",
      "difficulty_level": 3,
      "difficulty_name": "Advanced",
      "topic": "Bayesian Inference – Posterior Analysis",
      "problem": "Suppose you are analyzing a clinical trial for a new drug. Historical data suggest that the probability of a patient responding to the standard treatment (prior probability) is about 0.2. In a new trial with the experimental drug, you observe that out of 30 patients, 10 respond positively. Assume a Beta(2,8) prior distribution for the response probability θ (where Beta(2,8) reflects the historical data). The likelihood is binomial: 10 successes out of 30 trials.",
      "tasks": [
        "1. Calculate the posterior distribution for θ.",
        "2. Find the posterior mean and posterior mode.",
        "3. Construct a 95% credible interval for θ (you may use a normal approximation or look up critical values for the Beta distribution)."
      ]
    },
    {
      "index": "F3.901",
      "category": "F",
      "difficulty_level": 3,
      "difficulty_name": "Advanced",
      "topic": "Logistic Regression – Binary Classification",
      "problem": "Suppose you are analyzing data from a medical study where the outcome is whether a patient develops a certain disease (yes/no), and the predictors are age (in years) and cholesterol level (mg/dL). The following logistic regression output is obtained:",
      "regression_table": {
        "variables": [
          {
            "variable": "Intercept",
            "coefficient": -6.5,
            "standard_error": 1.2,
            "p_value": "<0.001"
          },
          {
            "variable": "Age",
            "coefficient": 0.04,
            "standard_error": 0.01,
            "p_value": "<0.001"
          },
          {
            "variable": "Cholesterol",
      "coefficient": 0.02,
      "standard_error": 0.005,
      "p_value": "<0.001"
     }
    ]
   },
   "tasks": [
    "1. Write the equation of the logistic regression model.",
    "2. For a patient aged 50 with a cholesterol level of 200 mg/dL, calculate the log-odds of developing the disease.",
    "3. Calculate the predicted probability of developing the disease for this patient.",
    "4. Interpret the coefficients for age and cholesterol."
   ]
  }
 ],
 "distribution_summary": {
  "by_category": {
   "A_Probability_Bayes": 19,
   "B_Exploratory_Data_Analysis": 1,
   "C_Confidence_Intervals": 6,
   "D_Parameter_Estimation": 4,
   "E_Hypothesis_Testing": 12,
   "F_Bayesian_Inference": 8
  },
  "by_difficulty": {
   "basic": 10,
   "intermediate": 23,
   "advanced": 17
  },
  "difficulty_distribution_by_category": {
   "A": {"basic": 2, "intermediate": 11, "advanced": 6},
   "B": {"basic": 0, "intermediate": 0, "advanced": 1},
   "C": {"basic": 3, "intermediate": 2, "advanced": 1},
   "D": {"basic": 1, "intermediate": 3, "advanced": 0},
   "E": {"basic": 2, "intermediate": 6, "advanced": 4},
   "F": {"basic": 1, "intermediate": 1, "advanced": 6}
  },
```

Statistics Solution Bank - Marker Guidance

Overview

This document contains 50 statistics questions with detailed solutions, organized using an A-F categorization system. Each question includes step-by-step solutions, conclusions, keywords, and academic sources.

Dataset Structure

Indexing System

Each question is labeled using the format: Letter.Level.Index (e.g., **A1.234**)

This unique identifier helps in tracking, filtering, and mapping questions across the broader dataset.

- **Letter**: Indicates the **topic category** (A–F)
- **Level (1, 2, 3)**: Indicates the **difficulty tier**
  - **1** = Basic: Foundational concepts, direct applications
  - **2** = Intermediate: Multi-step reasoning, applied contexts
  - **3** = Advanced: Complex statistical inference or modeling
- **Index**: Represents a simulated **entry number** in a 2000-question dataset

Example:

A2.187 refers to a **Probability & Bayes** question, at **Intermediate** difficulty, and is the 187th entry in the master bank.

Category Framework

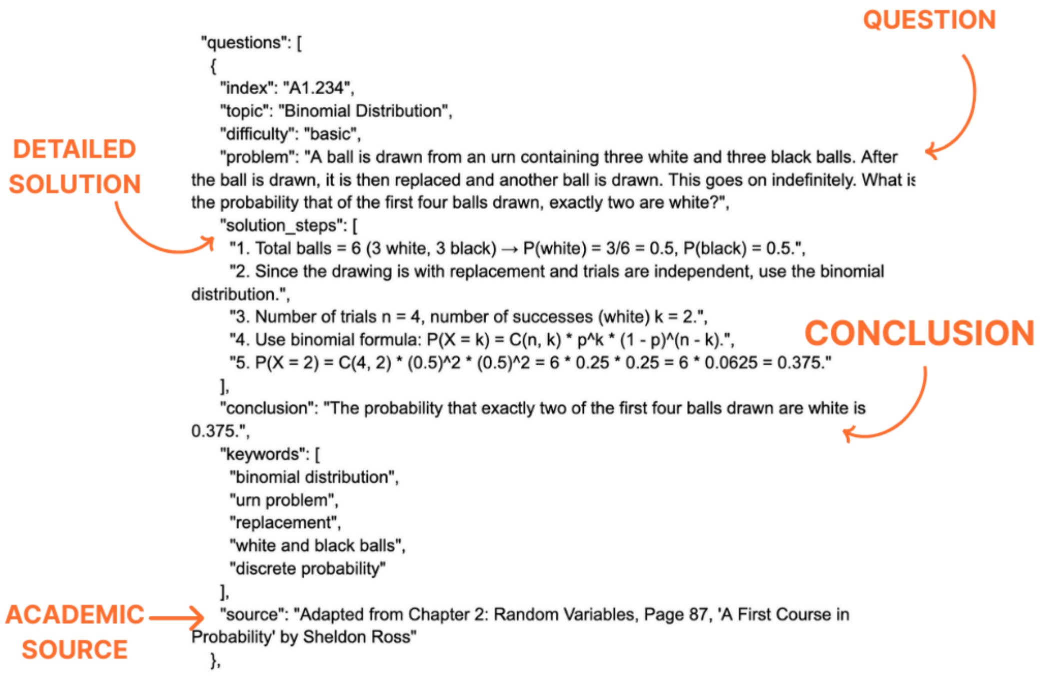
Code	Category	Description	Question Count
A	Probability & Bayes	Binomial, Poisson, Normal distributions, conditional probability, Markov chains	19 (38%)
B	Exploratory Data Analysis	Data visualization, descriptive statistics, gender/demographic analysis	1 (2%)
C	Confidence Intervals	Population means, proportions, sample size calculations	6 (12%)
D	Parameter Estimation	Sampling distributions, linear regression, Central Limit Theorem	4 (8%)
E	Hypothesis Testing	ANOVA, t-tests, z-tests, proportions, non-parametric methods	12 (24%)
F	Bayesian Inference	Bayes' theorem, posterior analysis, logistic regression	8 (16%)

Difficulty Levels

- **Basic (1)**: Introductory concepts, fundamental applications - 10 questions (20%)
- **Intermediate (2)**: Multi-step problems, moderate complexity - 23 questions (46%)
- **Advanced (3)**: Complex scenarios, sophisticated analysis - 17 questions (34%)

How to Read This Structured Solution Format

The JSON layout below is designed to present complex statistical problems in a standardized, machine-readable format.



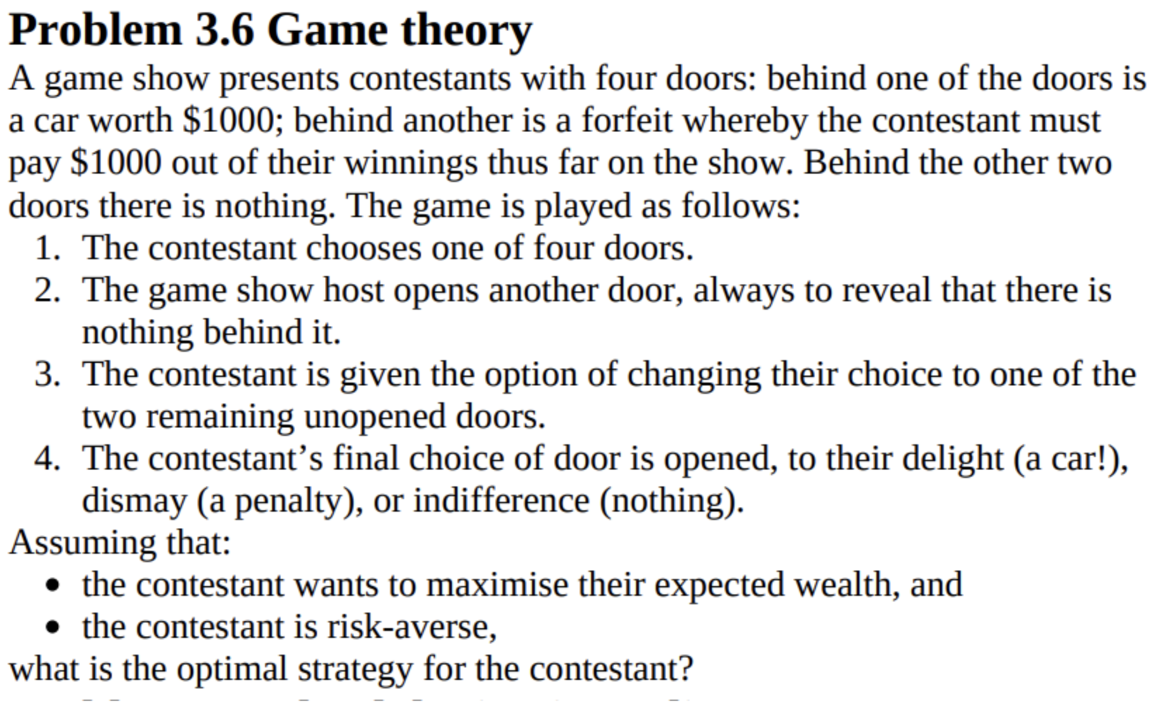
```
 "questions": [
  {
   "index": "A1.234",
   "topic": "Binomial Distribution",
   "difficulty": "basic",
   "problem": "A ball is drawn from an urn containing three white and three black balls. After
the ball is drawn, it is then replaced and another ball is drawn. This goes on indefinitely. What is
the probability that of the first four balls drawn, exactly two are white?",
   "solution_steps": [
    "1. Total balls = 6 (3 white, 3 black) → P(white) = 3/6 = 0.5, P(black) = 0.5.",
    "2. Since the drawing is with replacement and trials are independent, use the binomial
distribution.",
    "3. Number of trials n = 4, number of successes (white) k = 2.",
    "4. Use binomial formula: P(X = k) = C(n, k) * p^k * (1 - p)^(n - k).",
    "5. P(X = 2) = C(4, 2) * (0.5)^2 * (0.5)^2 = 6 * 0.25 * 0.25 = 6 * 0.0625 = 0.375."
   ],
   "conclusion": "The probability that exactly two of the first four balls drawn are white is
0.375.",
   "keywords": [
    "binomial distribution",
    "urn problem",
    "replacement",
    "white and black balls",
    "discrete probability"
   ],
   "source": "Adapted from Chapter 2: Random Variables, Page 87, 'A First Course in
Probability' by Sheldon Ross"
  },
```

Refer to the orange annotations in the visual to understand how each section contributes to a complete and traceable explanation.

If you encounter issues with specific questions or need clarification on marking criteria, please note the question index (e.g., A2.456) for reference.

📩 **Contact**: c.a.nagarkar@leeds.ac.uk

```
{
 "dataset_info": {
  "total_questions": 50,
  "indexing_system": "Letter-Level-Number format (e.g., A1.234)",
  "sampled_from": "2000 question dataset ",
  "categories": {
   "A": "Probability & Bayes",
   "B": "Exploratory Data Analysis",
   "C": "Confidence Intervals",
   "D": "Parameter Estimation",
   "E": "Hypothesis Testing",
   "F": "Bayesian Inference"
  },
  "difficulty_levels": {
   "1": "Basic",
   "2": "Intermediate",
   "3": "Advanced"
  }
 },
  "questions": [
    {
      "index": "A1.234",
      "topic": "Binomial Distribution",
      "difficulty": "basic",
      "problem": "A ball is drawn from an urn containing three white and three black balls. After
the ball is drawn, it is then replaced and another ball is drawn. This goes on indefinitely. What is
the probability that of the first four balls drawn, exactly two are white?",
      "solution_steps": [
        "1. Total balls = 6 (3 white, 3 black) → P(white) = 3/6 = 0.5, P(black) = 0.5.",
        "2. Since the drawing is with replacement and trials are independent, use the binomial
distribution.",
        "3. Number of trials n = 4, number of successes (white) k = 2.",
        "4. Use binomial formula: P(X = k) = C(n, k) * p^k * (1 - p)^(n - k).",
        "5. P(X = 2) = C(4, 2) * (0.5)^2 * (0.5)^2 = 6 * 0.25 * 0.25 = 6 * 0.0625 = 0.375."
      ],
      "conclusion": "The probability that exactly two of the first four balls drawn are white is
0.375.",
      "keywords": [
        "binomial distribution",
        "urn problem",
        "replacement",
        "white and black balls",
        "discrete probability"
      ],
      "source": "Adapted from Chapter 2: Random Variables, Page 87, 'A First Course in
Probability' by Sheldon Ross"
    },
```



Problem 3.6 Game theory

A game show presents contestants with four doors: behind one of the doors is a car worth $1000; behind another is a forfeit whereby the contestant must pay $1000 out of their winnings thus far on the show. Behind the other two doors there is nothing. The game is played as follows:

1. The contestant chooses one of four doors.
2. The game show host opens another door, always to reveal that there is nothing behind it.
3. The contestant is given the option of changing their choice to one of the two remaining unopened doors.
4. The contestant's final choice of door is opened, to their delight (a car!), dismay (a penalty), or indifference (nothing).

Assuming that:

- the contestant wants to maximise their expected wealth, and
- the contestant is risk-averse,

what is the optimal strategy for the contestant?

```
{
    "index": "A1.567",
    "topic": "Probability",
    "difficulty": "basic",
    "problem": "A game show has a prize with a 5% chance of being won. If a contestant plays 7 games, what is the probability of winning the prize at least once?",
    "solution_steps": [
      "Step 1: Use the complement rule: P(at least one win) = 1 - P(no wins).",
      "Step 2: Calculate the probability of not winning in one game: 1 - 0.05 = 0.95.",
      "Step 3: Compute the probability of not winning in 7 games: 0.95^7.",
      "Step 4: Calculate: 0.95^7 ≈ 0.698.",
      "Step 5: Find the probability of winning at least once: 1 - 0.698 = 0.302.",
      "Step 6: Thus, the probability of winning the prize at least once in 7 games is approximately 0.302."
    ],
    "conclusion": "The probability of winning the prize at least once in 7 games is approximately 0.302.",
    "keywords": [
      "probability",
      "complement rule",
      "game show",
      "at least one"
    ],
    "source": "Based on similar example from: Wyzant – 'Probability of winning at least 1 game out of 5 games' by Moonerah L."
    },
```

{
"index": "A2.123",
"topic": "Conditional Probability",
"difficulty": "intermediate",
"problem": "Among a group of 200 students, 137 students are enrolled in a mathematics class, 50 in a history class, and 124 in a music class. Additionally, 33 students are enrolled in both mathematics and history, 29 in both history and music, and 92 in both mathematics and music. Finally, 18 students are enrolled in all three classes. What is the probability that a randomly selected student is enrolled in at least one of the three classes?",
"solution_steps": [
"Let A_1 = enrolled in math, A_2 = enrolled in history, A_3 = enrolled in music.",
"Use the inclusion-exclusion principle:",
"$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_2 \cap A_3) - P(A_1 \cap A_3) + P(A_1 \cap A_2 \cap A_3)$",
"Substitute the given probabilities (out of 200 students):",
"$P(A_1 \cup A_2 \cup A_3) = (137 + 50 + 124 - 33 - 29 - 92 + 18) / 200 = 175 / 200 = 0.875$"
],
"conclusion": "The probability that a randomly selected student is enrolled in at least one of the three classes is 0.875 or 7/8.",
"keywords": [
"inclusion-exclusion",
"set theory",
"union of events",
"probability of overlapping events"
],
"source": "Probability and Statistics (Morris H. DeGroot, Mark J. Schervish), Chapter 1: 1.10 The Probability of a Union of Events, Example 1.10.1, Page 47"
},

```
 {
    "index": "A2.445",
    "topic": "Binomial Distribution",
    "difficulty": "intermediate",
    "problem": "Approximately 15% of the US population smokes cigarettes. A local government believed their community had a lower smoker rate and commissioned a survey of 400 randomly selected individuals. The survey found that only 42 of the 400 participants smoke cigarettes. If the true proportion of smokers in the community was really 15%, what is the probability of observing 42 or fewer smokers in a sample of 400 people?",
    "solution_steps": [
      "We are asked to compute P(k ≤ 42) where k is the number of smokers in a sample of n = 400 and the probability of success (p) is 0.15.",
      "This is the cumulative probability for k = 0, 1, 2, ..., 42 in a Binomial(n=400, p=0.15) distribution.",
      "Using binomial probability computations (or appropriate statistical software), we find:",
      "P(k ≤ 42) = 0.0054"
    ],
    "conclusion": "The probability of observing 42 or fewer smokers out of 400 when p = 0.15 is approximately 0.0054.",
    "keywords": [
      "binomial distribution",
      "cumulative probability",
      "sample proportion",
      "normal approximation"
    ],
    "source": "OpenIntro Statistics, 4th Edition (2019), Example 4.38, Chapter 4, Page 155"
  },
  {
    "index": "A2.678",
    "topic": "Probability",
    "difficulty": "intermediate",
    "problem": "A factory produces 1000 light bulbs, of which 40 are defective. If 5 bulbs are randomly selected without replacement, what is the probability that exactly 2 of them are defective?",
    "solution_steps": [
      "Step 1: Calculate the total number of ways to choose 5 bulbs from 1000: C(1000, 5).",
      "Step 2: Calculate the number of ways to choose exactly 2 defective bulbs from 40: C(40, 2).",
      "Step 3: Calculate the number of ways to choose 3 non-defective bulbs from 960: C(960, 3).",
      "Step 4: Compute the number of favorable outcomes: C(40, 2) * C(960, 3).",
      "Step 5: Divide by the total number of possible outcomes: (C(40, 2) * C(960, 3)) / C(1000, 5).",
      "Step 6: Using the binomial coefficient calculations: C(1000, 5) = 2,424,789,642, C(40, 2) = 780, C(960, 3) = 1,297,060, C(1000, 5) = 2,424,789,642.",
      "Step 7: The probability is: (780 * 1,297,060) / 2,424,789,642 ≈ 0.218."
    ],
    "conclusion": "The probability that exactly 2 of the 5 selected bulbs are defective is approximately 0.218.",
    "keywords": [
      "hypergeometric distribution",
      "defective bulbs",
      "without replacement",
      "combinatorics"
    ],
    "source": "STAT 414: Introduction to Probability Theory, Lesson 7.4 - Hypergeometric Distribution, Example 7-7, Penn State University"
  },
  {
    "index": "A2.789",
    "topic": "Binomial Distribution",
    "difficulty": "intermediate",
    "problem": "At Selitall Supermarket, 60% of customers pay by credit card. Find the probability that in a randomly selected sample of ten customers, (a) exactly two pay by credit card, (b) more than seven pay by credit card.",
    "solution_steps": [
      "Let X be the number of customers who pay by credit card in a sample of 10.",
      "Success: 'pays by credit card' → p = 0.6, q = 0.4, n = 10.",
      "Then X ~ B(10, 0.6).",
      "(a) To find P(X = 2):",
      "Use binomial formula: P(X = 2) = 10C2 × (0.4)^8 × (0.6)^2",
      "= 45 × 0.4^8 × 0.6^2 ≈ 0.011 (2 s.f.)",
      "(b) To find P(X > 7):",
      "P(X > 7) = P(X = 8) + P(X = 9) + P(X = 10)",
      "= 10C8 × 0.4^2 × 0.6^8 + 10C9 × 0.4^1 × 0.6^9 + 10C10 × 0.6^10",
      "= 45 × 0.4^2 × 0.6^8 + 10 × 0.4 × 0.6^9 + 0.6^10",
      "≈ 0.17 (2 s.f.)"
    ],
    "conclusion": "The probability that exactly two out of ten customers pay by credit card is approximately 0.011. The probability that more than seven customers pay by credit card is approximately 0.17.",
    "keywords": [
      "binomial probability",
      "credit card",
      "sampling",
      "binomial distribution",
      "greater than probability"
    ],
    "source": "A Concise Course in A-Level Statistics with worked examples – 26 Jun. 2013, Example 5.7, Page 280"
  },
  {
    "index": "A2.345",
    "topic": "Poisson Distribution",
    "difficulty": "intermediate",
    "problem": "Two identical racing cars are being tested on a circuit. For each car, the number of mechanical breakdowns can be modelled by a Poisson distribution with a mean of one breakdown in 100 laps. If a car breaks down, it is attended and continues on the circuit. The first car is tested for 20 laps and the second car for 40 laps. Find the probability that the service team is called out to attend to breakdowns: (a) once, (b) more than twice.",
    "solution_steps": [
      "Let X be the number of breakdowns of the first car: X ~ Poisson(0.2) since 20 laps × 1/100 = 0.2.",
      "Let Y be the number of breakdowns of the second car: Y ~ Poisson(0.4) since 40 laps × 1/100 = 0.4.",
      "Let T = X + Y be the total number of breakdowns, so T ~ Poisson(0.6).",
      "(a) P(T = 1) = 0.6 × e^(-0.6) ≈ 0.329 (3 d.p.)",
      "(b) P(T > 2) = 1 - [P(T = 0) + P(T = 1) + P(T = 2)]",
      "     = 1 - e^(-0.6) × (1 + 0.6 + 0.6^2 / 2!)",
      "     = 1 - e^(-0.6) × (1 + 0.6 + 0.18) = 1 - e^(-0.6) × 1.78 ≈ 0.023 (3 d.p.)"
    ],
    "conclusion": "The probability of one breakdown is approximately 0.329. The probability of more than two breakdowns is approximately 0.023.",
    "keywords": [
      "Poisson distribution",
      "total probability",
      "breakdowns",
      "racing cars",
      "discrete distributions"
    ],
    "source": "A Concise Course in Advanced Level Statistics with worked examples – 26 Jun. 2013, Example 5.25"
  },
  {
    "index": "A2.456",
    "topic": "Normal Distribution",
    "difficulty": "intermediate",
    "problem": "Lengths of metal strips produced by a machine are normally distributed with a mean length of 150 cm and a standard deviation of 10 cm. Find the probability that the length of a randomly selected strip is: (a) shorter than 165 cm, (b) within 5 cm of the mean.",
    "solution_steps": [
      "Let X be the length in cm of a metal strip.",
      "Given: μ = 150, σ = 10. So X ~ N(150, 10²).",
      "(a) Find P(X < 165):",
      "Standardize: Z = (X - μ) / σ = (165 - 150) / 10 = 1.5",
      "Look up Z = 1.5 in standard normal table: Φ(1.5) ≈ 0.9332",
      "So, P(X < 165) = 0.9332 ≈ 0.93 (2 s.f.)",
      "(b) Find P(145 < X < 155):",
      "Standardize lower bound: Z₁ = (145 - 150) / 10 = -0.5",
      "Standardize upper bound: Z₂ = (155 - 150) / 10 = 0.5",
      "Look up values: Φ(0.5) = 0.6915, Φ(-0.5) = 1 - Φ(0.5) = 0.3085",
      "P(145 < X < 155) = Φ(0.5) - Φ(-0.5) = 0.6915 - 0.3085 = 0.383 ≈ 0.38 (2 s.f.)"
    ],
    "conclusion": "The probability that the length is shorter than 165 cm is approximately 0.93. The probability that the length is within 5 cm of the mean is approximately 0.38.",
    "keywords": [
      "normal distribution",
      "standardization",
      "Z-score",
      "probability",
      "mean",
      "standard deviation"
    ],
    "source": "A Concise Course in A-Level Statistics with worked examples – 26 Jun. 2013, Example 7.4, Page 368"
  },
  {
    "index": "A2.567",
      "topic": "Probability",
  "difficulty": "advanced",
  "problem": "A gambler plays a fair game by flipping a coin. Heads wins $1, tails loses $1. The gambler starts with $50 and stops playing when reaching $0 (ruin) or $200. What is the probability of reaching $200 before going broke?",
  "solution_steps": [
    "Step 1: Define the winning probability function y(n) = P(200 | n), where n is the current amount of money.",
    "Step 2: Use boundary conditions: y(0) = 0 (ruin) and y(200) = 1 (goal reached).",
    "Step 3: Apply the recurrence relation for a fair coin: y(n) = 0.5 * y(n + 1) + 0.5 * y(n - 1).",
    "Step 4: This implies that y(n + 1) - y(n) = y(n) - y(n - 1), so the graph of y(n) is linear.",
    "Step 5: Given y(0) = 0 and y(200) = 1, the linear solution is y(n) = n / 200.",
    "Step 6: Substitute n = 50 to find the probability: y(50) = 50 / 200 = 0.25."
  ],
  "conclusion": "The probability of reaching $200 before going broke when starting with $50 is 0.25.",
  "keywords": [
    "gambler's ruin",
    "probability",
    "coin toss",
    "recurrence relation",
    "stochastic process"
  ],
 "source": "Example 4.8, 'Gambler’s Ruin', Chapter 4 – Probability, University of Connecticut OER: https://probability.oer.math.uconn.edu/wp-content/uploads/sites/2187/2018/01/prob3160ch4.pdf"
},
```

```
  {
    "index": "A2.011",
    "topic": "Discrete Probability",
```

"difficulty": "intermediate",
"problem": "Five people have just won a $100 prize and are deciding how to divide the $100 up between them. Assume that whole dollars are used, not cents. Also, for example, giving $50 to the first person and $10 to the second is different from vice versa. (a) How many ways are there to divide up the $100, such that each person gets at least $10?",
"solution_steps": [
"1. Since each person must get at least $10, we first give $10 to each person: 5 × $10 = $50.",
"2. We are left with $100 - $50 = $50 to distribute freely (each person can now get $0 or more).",
"3. This is a classic stars and bars problem: distributing 50 indistinguishable dollars among 5 distinguishable people, with no minimum requirement.",
"4. The number of integer solutions to $x_1 + x_2 + x_3 + x_4 + x_5 = 50$ where $x_i \geq 0$ is given by the formula: C(n + k - 1, k), where k = 50 and n = 5.",
"5. So, the number of solutions is: C(50 + 5 - 1, 50) = C(54, 4)."
],
"conclusion": "The number of ways to divide the $100 such that each person gets at least $10 is C(54, 4).",
"keywords": [
"stars and bars",
"integer partition",
"combinatorics",
"minimum allocation",
"Bose-Einstein counting"
],
"source": "Introduction to Probability by Joseph K. Blitzstein and Jessica Hwang, Chapter 4, Question 77(a), Page 41"
},

```
{
  "index": "A2.012",
  "topic": "Conditional Probability",
  "difficulty": "intermediate",
  "problem": "A game show presents contestants with four doors: behind one is a car worth $1000, behind another is a forfeit costing -$1000, and behind the remaining two doors is nothing. The contestant picks one door. The host then opens one of the remaining doors to show it is empty. The contestant may stick with their choice or switch to one of the other two unopened doors. What is the optimal strategy if the contestant wants to maximise expected wealth and is risk-averse?",
  "solution_steps": [
    "Step 1: Initial probabilities before any door is opened: P(car) = 1/4 = 0.25, P(forfeit) = 1/4 = 0.25, P(nothing) = 2/4 = 0.50",
    "Step 2: Expected value of original choice (stick strategy): EV_stick = 0.25 × 1000 + 0.25 × (-1000) + 0.50 × 0 = 250 - 250 + 0 = $0",
    "Step 3: Host opens one of the remaining doors to show it is empty (always possible since there are two empty doors among the three not chosen). After this action, 2 unopened doors remain besides the original pick.",
    "Step 4: Analyze switching strategy by considering what the contestant originally picked. If contestant initially picked a 'nothing' door (probability = 0.5): Among the other 3 doors, host removes the other 'nothing' door, leaving 'car' and 'forfeit'. Switching means randomly choosing between these: P(car|switch, initial nothing) = 0.5, P(forfeit|switch, initial nothing) = 0.5",
    "Step 5: If contestant originally picked the car (probability = 0.25): Among the other 3 doors are 1 forfeit and 2 nothing. Host removes 1 nothing door, leaving forfeit and nothing. Switching gives: P(forfeit|switch, initial car) = 0.5, P(nothing|switch, initial car) = 0.5",
    "Step 6: If contestant originally picked the forfeit (probability = 0.25): Among the other 3 doors are 1 car and 2 nothing. Host removes 1 nothing door, leaving car and nothing. Switching gives: P(car|switch, initial forfeit) = 0.5, P(nothing|switch, initial forfeit) = 0.5",
    "Step 7: Calculate overall probabilities under switching strategy: P(car) = 0.5 × 0.5 + 0.25 × 0 + 0.25 × 0.5 = 0.25 + 0 + 0.125 = 3/8. P(forfeit) = 0.5 × 0.5 + 0.25 × 0.5 + 0.25 × 0 = 0.25 + 0.125 + 0 = 3/8. P(nothing) = 0.5 × 0 + 0.25 × 0.5 + 0.25 × 0.5 = 0 + 0.125 + 0.125 = 1/4",
    "Step 8: Expected value of switching: EV_switch = (3/8) × 1000 + (3/8) × (-1000) + (1/4) × 0 = 375 - 375 + 0 = $0",
    "Step 9: Calculate variance for both strategies. Variance_stick = (1/4)(1000)² + (1/4)(-1000)² + (1/2)(0)² = 250,000 + 250,000 + 0 = 500,000. Variance_switch = (3/8)(1000)² + (3/8)(-1000)² + (1/4)(0)² = 375,000 + 375,000 + 0 = 750,000",
    "Step 10: Since both strategies have identical expected values ($0), the risk-averse contestant should choose the strategy with lower variance. Sticking has variance of 500,000 while switching has variance of 750,000, making sticking the optimal choice for risk-averse players."
  ],
  "conclusion": "Although both strategies yield an expected value of $0, the optimal strategy for a risk-averse contestant is to **stick** with their original choice. While switching increases the
probability of winning the car from 25% to 37.5%, it also increases the probability of the forfeit
from 25% to 37.5%, resulting in higher variance (750,000 vs 500,000). When expected values
are equal, variance becomes the determining factor for risk-averse decision-making, making
sticking with the original choice the optimal strategy.",
  "keywords": [
    "expected value",
    "conditional probability",
    "risk aversion",
    "variance",
    "decision theory",
    "game theory",
    "Monty Hall variation"
  ],
  "source": "Custom analysis of a 4-door Monty Hall-style game"
}
```

```
 {
    "index": "A2.890",
    "topic": "Conditional Probability – Boy or Girl Paradox",
    "difficulty": "intermediate",
    "problem": "Mr Bayes has two children. The older child is a girl. What is the probability that both children are girls?",
    "solution_steps": [
      "Step 1: List all possible gender combinations of two children: (Boy, Boy), (Boy, Girl), (Girl, Boy), (Girl, Girl).",
      "Step 2: Given that the older child is a girl, eliminate combinations where the older child is a boy.",
      "Step 3: The remaining valid combinations are: (Girl, Boy) and (Girl, Girl).",
      "Step 4: Out of these 2 equally likely combinations, only (Girl, Girl) has both children as girls."
    ],
    "conclusion": "The probability that both children are girls given the older child is a girl is 1/2.",
    "keywords": [
      "conditional probability",
      "boy or girl paradox",
      "Bayes",
      "children combinations"
    ],
    "source": "A Student's Guide to Bayesian Statistics by Ben Lambert, Problem 3.4.1"
  },
  {
    "index": "A2.901",
    "topic": "Conditional Probability – Boy or Girl Paradox",
    "difficulty": "intermediate",
    "problem": "Mr Laplace has two children. At least one of the children is a girl. What is the probability that both children are girls?",
    "solution_steps": [
      "Step 1: List all possible gender combinations: (Boy, Boy), (Boy, Girl), (Girl, Boy), (Girl, Girl).",
      "Step 2: Given that at least one child is a girl, eliminate the (Boy, Boy) case.",
      "Step 3: Remaining valid combinations: (Boy, Girl), (Girl, Boy), (Girl, Girl).",
      "Step 4: Out of these 3 equally likely combinations, only (Girl, Girl) has both children as girls."
    ],
    "conclusion": "The probability that both children are girls given at least one is a girl is 1/3.",
    "keywords": [
      "conditional probability",
      "boy or girl paradox",
      "Laplace",
      "children combinations"
    ],
    "source": "A Student's Guide to Bayesian Statistics by Ben Lambert, Problem 3.4.2"
  },
 {
    "index": "A3.123",
    "topic": "Probability",
    "difficulty": "advanced",
    "problem": "We roll a fair four-sided die. If the result is 1 or 2, we roll once more but otherwise, we stop. What is the probability that the sum total of our rolls is at least 4?",
    "solution_steps": [
      "1. Let $A_1$ be the event that the first roll is 1.",
      "2. Let $A_2$ be the event that the first roll is 2.",
      "3. Let $A_3$ be the event that the first roll is 3.",
      "4. Let $A_4$ be the event that the first roll is 4.",
      "5. For $A_1$: second roll needed. Probability that second roll is 3 or 4 (to make total ≥ 4) = 2/4 = 1/2.",
      "6. For $A_2$: need second roll ≥ 2 to make sum ≥ 4. Probability = 3/4.",
      "7. For $A_3$: no reroll, total is 3 < 4 ⇒ probability = 0.",
      "8. For $A_4$: total is already 4 ⇒ probability = 1.",
      "9. Use Total Probability Theorem:",
      "   $P(B) = P(A_1)·P(B|A_1) + P(A_2)·P(B|A_2) + P(A_3)·P(B|A_3) + P(A_4)·P(B|A_4)$",
      "        = 1/4·1/2 + 1/4·3/4 + 1/4·0 + 1/4·1",
      "        = 1/8 + 3/16 + 0 + 1/4 = 9/16"
    ],
    "conclusion": "The probability that the sum total of the rolls is at least 4 is 9/16.",
    "keywords": [
      "sequential probability",
      "law of total probability",
      "conditional probability",
      "discrete distribution",
      "multi-stage experiment"
    ],
    "source": "Example 1.13, Page 28, *Introduction to Probability* by Dimitri P. Bertsekas and John N. Tsitsiklis (2nd Edition, 2008)"
  },
 {
    "index": "A3.234",
    "topic": "Probability",
    "difficulty": "advanced",
    "problem": "A store has 20% off sales on items. If a customer buys 8 items, what is the
probability that at least 3 of the items are on sale?",
    "solution_steps": [
      "Step 1: Use the complement rule to find the probability of fewer than 3 items on sale: P(X
< 3).",
      "Step 2: Calculate the probability of 0, 1, and 2 items on sale using the binomial formula:
P(X = k) = C(n, k) * p^k * (1-p)^(n-k).",
      "Step 3: Compute: P(X = 0) = C(8, 0) * (0.20)^0 * (0.80)^8 ≈ 0.167, P(X = 1) = C(8, 1) *
(0.20)^1 * (0.80)^7 ≈ 0.335, P(X = 2) = C(8, 2) * (0.20)^2 * (0.80)^6 ≈ 0.285.",
      "Step 4: Sum these probabilities: P(X < 3) ≈ 0.167 + 0.335 + 0.285 = 0.787.",
      "Step 5: Calculate the probability of at least 3 items on sale: 1 - P(X < 3) ≈ 1 - 0.787 =
0.213."
    ],
    "conclusion": "The probability that at least 3 out of 8 items are on sale is approximately
0.213.",
    "keywords": [
      "binomial distribution",
      "complement rule",
      "probability calculation",
      "at least probability"
    ],
    "source": "Based on a similar example from: Statology – 'Probability of at Least Three'"
  },
    {
      "index": "A3.345",
      "topic": "Probability",
      "difficulty": "advanced",
{
  "problem": "In a population of 200 people, 30 have a certain disease. If 10 people are randomly selected, what is the probability that exactly 4 of them have the disease?",
  "solution_steps": [
    "Step 1: Identify that this follows a hypergeometric distribution since we're sampling without replacement from a finite population. Use the formula: P(X = k) = (C(K, k) × C(N-K, n-k)) / C(N, n)",
    "Step 2: Define the parameters: N = 200 (total population), K = 30 (people with disease), n = 10 (sample size), k = 4 (exactly 4 with disease). So P(X = 4) = (C(30, 4) × C(170, 6)) / C(200, 10)",
    "Step 3: Calculate C(30, 4): C(30, 4) = 30!/(4! × 26!) = (30 × 29 × 28 × 27)/(4 × 3 × 2 × 1) = 27,405",
    "Step 4: Calculate C(170, 6): C(170, 6) = 170!/(6! × 164!) = (170 × 169 × 168 × 167 × 166 × 165)/(6 × 5 × 4 × 3 × 2 × 1) = 30,663,442,810",
    "Step 5: Calculate C(200, 10): C(200, 10) = 200!/(10! × 190!) = 22,451,004,309,013,280",
    "Step 6: Calculate the probability: P(X = 4) = (27,405 × 30,663,442,810) / 22,451,004,309,013,280 = 840,331,650,208,050 / 22,451,004,309,013,280 ≈ 0.0374",
    "Step 7: Therefore, the probability that exactly 4 out of 10 randomly selected people have the disease is approximately 0.037 or 3.74%"
  ]
}
 "conclusion": "The probability that exactly 4 out of 10 selected people have the disease is approximately 0.037 or 3.74%. This relatively low probability makes sense given that only 15% of the population has the disease, so getting 40% diseased individuals in a small sample is unlikely.",
      "keywords": [
        "hypergeometric distribution",
        "disease probability",
        "sampling without replacement",
        "combinatorics"
      ],
      "source": "Understanding Advanced Statistical Methods by Westfall & Henning"
    },
  {
    "index": "A3.456",
    "topic": "Bias and Logistic Thinking",
    "difficulty": "advanced",
    "problem": "Let's consider a sex-imbalanced company that consists of 20% women and 80% men. Suppose that the company is very large, consisting of perhaps 20,000 employees. Suppose when someone goes up for promotion at this company, 5 of their colleagues are randomly chosen to provide feedback on their work. Now let's imagine that 10% of the people in the company are prejudiced against the other sex. That is, 10% of men are prejudiced against women, and similarly, 10% of women are prejudiced against men. Who is discriminated against more at the company, men or women?",
    "solution_steps": [
      "For 100 men going for promotion, 5 × 100 = 500 colleagues will review them. Since the company has 20% women, approximately 100 of the 500 reviewers are women.",
      "Of these 100 women, 10% = 10 are biased against men, resulting in 10 biased reviews out of 500 → 2% discrimination rate.",
      "For 100 women going for promotion, again 500 colleagues provide feedback. With 80% men in the company, 400 of the 500 reviewers are men.",
      "Of these 400 men, 10% = 40 are biased against women, leading to 40 biased reviews out of 500 → 8% discrimination rate."
    ],
    "conclusion": "Women experience higher discrimination (8%) compared to men (2%) in this scenario, despite equal bias percentages among colleagues.",
    "keywords": [
      "discrimination",
      "bias",
      "promotion",
      "gender imbalance",
      "logistic reasoning"
    ],
    "source": "OpenIntro Statistics, 4th Edition (2019), Example 9.35, Chapter 9, Page 379"
  },
 {
   "index": "A3.567",
   "topic": "Probability Distributions",
   "difficulty": "advanced",
   "problem": "Heights of 10 year olds, regardless of gender, closely follow a normal distribution with mean 55 inches and standard deviation 6 inches. (a) What fraction of 10 year olds are taller than 76 inches? (b) If there are 2,000 10 year olds entering Six Flags Magic Mountain in a single day, then compute the expected number of 10 year olds who are at least 76 inches tall. (Assume independence.) (c) Using the binomial distribution, compute the probability that 0 of the 2,000 10 year olds will be at least 76 inches tall. (d) The number of 10 year olds who enter Six Flags Magic Mountain and are at least 76 inches tall in a given day follows a Poisson distribution with mean equal to the value found in part (b). Use the Poisson distribution to identify the probability that no 10 year old will enter the park who is 76 inches or taller.",
   "solution_steps": [
     "(a) Use the standard normal distribution: z = (76 - 55) / 6 = 3.5. Then P(Z > 3.5) ≈ 0.00023. So, approximately 0.023% of 10-year-olds are taller than 76 inches.",
     "(b) Expected number = 2000 × 0.00023 ≈ 0.46 children.",
     "(c) Use Binomial distribution: n = 2000, p = 0.00023. P(X = 0) = (1 - 0.00023)^2000 ≈ e^(-0.46) ≈ 0.631.",
     "(d) Use Poisson distribution with λ = 0.46. P(X = 0) = e^(-0.46) ≈ 0.631."
   ],
   "conclusion": "The expected number of children ≥76 inches is less than 1, and the probability of observing 0 such children out of 2,000 is approximately 63%.",
   "keywords": [
     "normal distribution",
     "z-score",
     "binomial distribution",
     "poisson distribution",
     "expected value",
     "tail probability"
   ],
   "source": "OpenIntro Statistics, 4th Edition (2019), Example 4.47, Chapter 4, Page 151"
 },
```

{
"index": "A3.678",
"topic": "Poisson Distribution with Uncertainty",
"difficulty": "advanced",
"problem": "The opponents of soccer team A are of two types: either they are a class 1 or a class 2 team. The number of goals team A scores against a class i opponent is a Poisson random variable with mean λ_i, where $\lambda_1 = 2$, $\lambda_2 = 3$. This weekend, the team has two games against teams they are not very familiar with. Assuming that the first team they play is a class 1 team with probability 0.6 and the second is, independently of the class of the first team, a class 1 team with probability 0.3, determine: (a) the expected number of goals team A will score this weekend; (b) the probability that team A will score a total of five goals.",
"solution_steps": [
"(a) Let X_1 and X_2 be the number of goals in the first and second games. Define $\lambda_1 = 2$ (class 1 team), $\lambda_2 = 3$ (class 2 team). $E[X_1] = 0.6 \times 2 + 0.4 \times 3 = 1.2 + 1.2 = 2.4$. $E[X_2] = 0.3 \times 2 + 0.7 \times 3 = 0.6 + 2.1 = 2.7$. E[Total Goals] = $E[X_1] + E[X_2] = 2.4 + 2.7 = 5.1$",
"(b) We need to compute $P(X_1 + X_2 = 5)$, where X_1 and X_2 are independent Poisson mixtures. Use total probability by conditioning on combinations of class types (there are 4 cases), then weight each by the probability of class combination. Case 1: Both class 1 → P = 0.6 × 0.3 = 0.18; λ_total = 2 + 2 = 4 → P(X=5) = e^-4 * 4^5 / 5! = 0.1563. Case 2: First class 1, second class 2 → P = 0.6 × 0.7 = 0.42; λ_total = 2 + 3 = 5 → P(X=5) = e^-5 * 5^5 / 5! = 0.1755. Case 3: First class 2, second class 1 → P = 0.4 × 0.3 = 0.12; λ_total = 3 + 2 = 5 → P(X=5) = 0.1755. Case 4: Both class 2 → P = 0.4 × 0.7 = 0.28; λ_total = 3 + 3 = 6 → P(X=5) = e^-6 * 6^5 / 5! = 0.1606. Weighted sum: 0.18×0.1563 + 0.42×0.1755 + 0.12×0.1755 + 0.28×0.1606 ≈ 0.0281 + 0.0737 + 0.0211 + 0.0450 = 0.168"
],
"conclusion": "The expected number of goals team A will score this weekend is 5.1. The probability that team A will score exactly 5 goals is approximately 0.168.",
"keywords": [
"Poisson distribution",
"mixture model",
"expected value",
"total probability",
"discrete distribution"

```
    ],
    "source": "Adapted from Sheldon Ross, A First Course in Probability, 9th Edition, Chapter 3: Conditional Probability and Conditional Expectation, Page 183"
  },



  {
    "index": "B3.789",
    "topic": "Gender and Degree Distributions",
    "difficulty": "advanced",
    "problem": "Using the provided counts (in thousands) of earned degrees in the U.S. by gender and degree type, perform an exploratory data analysis to understand trends and imbalances. Specifically, analyze: (1) overall gender distribution, (2) the distribution of degrees within each gender, (3) the conditional probability of selecting a male given a Master's degree, (4) gender gap index for each degree type, and (5) the most gender-skewed degree type. Data: Bachelor's: Female 616, Male 529; Master's: Female 194, Male 171; Professional: Female 30, Male 44; Doctorate: Female 16, Male 26.",
    "solution_steps": [
      "Step 1: Compute total number of degree recipients: Total = 616 + 529 + 194 + 171 + 30 + 44 + 16 + 26 = 1626 (in thousands).",
      "Step 2: Calculate total recipients by gender: Total Female = 616 + 194 + 30 + 16 = 856, Total Male = 529 + 171 + 44 + 26 = 770",
      "Step 3: Compute overall gender distribution: P(Female) = 856 / 1626 ≈ 0.5264 (52.64%), P(Male) = 770 / 1626 ≈ 0.4736 (47.36%)",
      "Step 4: Calculate the proportion of each degree type within each gender: For Females: Bachelor's = 616 / 856 ≈ 0.7196 (71.96%), Master's = 194 / 856 ≈ 0.2266 (22.66%), Professional = 30 / 856 ≈ 0.0350 (3.5%), Doctorate = 16 / 856 ≈ 0.0187 (1.87%). For Males: Bachelor's = 529 / 770 ≈ 0.6870 (68.70%), Master's = 171 / 770 ≈ 0.2221 (22.21%), Professional = 44 / 770 ≈ 0.0571 (5.71%), Doctorate = 26 / 770 ≈ 0.0338 (3.38%)",
      "Step 5: Calculate conditional probability: P(Male | Master's) = 171 / (171 + 194) = 171 / 365 ≈ 0.4685 (46.85%)",
      "Step 6: Compute the Gender Gap Index (GGI) for each degree type: GGI = (Female - Male) / (Female + Male). Bachelor's: (616 - 529) / (616 + 529) ≈ 0.076 (Female-dominant), Master's: (194 - 171) / (194 + 171) ≈ 0.063 (Female-dominant), Professional: (30 - 44) / (30 + 44) ≈ -0.1892 (Male-dominant), Doctorate: (16 - 26) / (16 + 26) ≈ -0.2381 (Male-dominant)",
      "Step 7: Identify the degree type with highest gender imbalance: Doctorate degree has the highest GGI magnitude (≈ 0.2381), indicating the strongest gender skew—toward males."
    ],
    "conclusion": "The majority of degree recipients are women (52.64%). However, gender dominance varies by degree level. While women dominate at Bachelor's and Master's levels, men are more prevalent in Professional and Doctorate degrees. The greatest gender disparity is observed in Doctorate degrees, heavily favoring men.",
    "keywords": [
      "EDA",
      "gender distribution",
      "degree analysis",
      "conditional probability",
      "gender gap index"
    ],
    "source": "OpenIntro Statistics Dataset Analysis"
  },
  {
    "index": "C1.234",
    "topic": "Confidence Interval",
    "difficulty": "basic",
    "problem": "A teacher records the test scores of 10 students with an average score of 78 and a standard deviation of 5. Calculate the 95% confidence interval for the mean test score.",
    "solution_steps": [
      "Step 1: Identify the sample mean (x̄) = 78, standard deviation (σ) = 5, and sample size (n) = 10.",
      "Step 2: Compute the standard error of the mean (SE) = σ / √n = 5 / √10 ≈ 1.581.",
      "Step 3: Find the t-value for a 95% confidence level with n-1 degrees of freedom (t ≈ 2.262).",
      "Step 4: Calculate the margin of error (ME) = t * SE = 2.262 * 1.581 ≈ 3.58.",
      "Step 5: Construct the confidence interval: Lower limit = x̄ - ME = 78 - 3.58 = 74.42, Upper limit = x̄ + ME = 78 + 3.58 = 81.58.",
      "Step 6: Thus, the 95% confidence interval is (74.42, 81.58)."
    ],
    "conclusion": "The 95% confidence interval for the mean test score is (74.42, 81.58).",
    "keywords": [
      "confidence interval",
      "t-distribution",
      "sample mean",
      "margin of error"
    ],
    "source": "Understanding Confidence Intervals - Statistical Guide"
  },
 {
    "index": "C1.456",
    "topic": "Confidence Interval",
    "difficulty": "basic",
    "problem": "A sample of 18 students has an average test score of 72 with a standard deviation of 8. Calculate the 95% confidence interval for the average test score.",
    "solution_steps": [
     "Step 1: Identify the sample mean (x̄) = 72, standard deviation (σ) = 8, and sample size (n) = 18.",
     "Step 2: Compute the standard error of the mean (SE) = σ / √n = 8 / √18 ≈ 1.887.",
     "Step 3: Find the t-value for a 95% confidence level with n-1 degrees of freedom (t ≈ 2.101).",
     "Step 4: Calculate the margin of error (ME) = t * SE = 2.101 * 1.887 ≈ 3.96.",
     "Step 5: Construct the confidence interval: Lower limit = x̄ - ME = 72 - 3.96 = 68.04, Upper limit = x̄ + ME = 72 + 3.96 = 75.96.",
     "Step 6: Thus, the 95% confidence interval is (68.04, 75.96)."
    ],
    "conclusion": "The 95% confidence interval for the average test score is (68.04, 75.96).",
    "keywords": [
     "confidence interval",
     "t-distribution",
     "sample statistics",
     "margin of error"
    ],
    "source": "Based on example from: University of West Georgia – Confidence Intervals Notes (https://www.westga.edu/academics/research/vrc/assets/docs/confidence_intervals_notes.pdf)""
  },
  {
    "index": "C1.789",
    "topic": "Population Proportion",
    "difficulty": "basic",
    "problem": "A market research firm is hired to estimate the proportion of adults in a large city who own smartphones. A random sample of 500 adult residents is surveyed, and 421 say they own smartphones. Using a 95% confidence level, compute a confidence interval for the true proportion of adult residents who own smartphones.",
    "solution_steps": [
      "Step 1: Define the known values. Sample size (n) = 500, Number of successes (x) = 421",
      "Step 2: Compute the sample proportion. p̂ = x / n = 421 / 500 = 0.842, q̂ = 1 - p̂ = 1 - 0.842 = 0.158",
      "Step 3: Determine the z-score for a 95% confidence level. α = 0.05 → α/2 = 0.025, z_{α/2} = 1.96 (from standard normal table or invNorm(0.975))",
      "Step 4: Calculate the margin of error (EBP). EBP = z * sqrt((p̂ * q̂) / n), EBP = 1.96 * sqrt((0.842 * 0.158) / 500) ≈ 0.032",
      "Step 5: Construct the confidence interval. Lower bound = p̂ - EBP = 0.842 - 0.032 = 0.810, Upper bound = p̂ + EBP = 0.842 + 0.032 = 0.874",
      "Step 6: Final confidence interval. (0.810, 0.874)"
    ],
    "conclusion": "We estimate with 95% confidence that the true proportion of adult residents in this city who own smartphones is between 81% and 87.4%.",
    "interpretation": "This means that if we took many such samples and built confidence intervals from each, approximately 95% of those intervals would contain the true population proportion.",
    "keywords": [
      "confidence interval",
      "population proportion",
      "binomial distribution",
      "margin of error",
      "z-score",
      "sample proportion"
      ],
      "source": "OpenStax: Introductory Statistics 2e, Section 8.3 (https://openstax.org/books/introductory-statistics-2e/pages/8-3-a-population-proportion )"
    },

    {
      "index": "C2.123",
      "topic": "Confidence Interval for a Mean",
      "difficulty": "intermediate",
      "problem": "The General Social Survey asked the question: 'For how many days during the past 30 days was your mental health, which includes stress, depression, and problems with emotions, not good?' Based on responses from 1,151 US residents, the survey reported a 95% confidence interval of 3.40 to 4.24 days in 2010. (a) Interpret this interval in context of the data. (b) What does '95% confident' mean? Explain in the context of the application. (c) Suppose the researchers think a 99% confidence level would be more appropriate for this interval. Will this new interval be smaller or wider than the 95% confidence interval? (d) If a new survey were to be done with 500 Americans, do you think the standard error of the estimate would be larger, smaller, or about the same?",
      "solution_steps": [
        "(a) We are 95% confident that the true average number of mentally unhealthy days in the past 30 days for US adults in 2010 was between 3.40 and 4.24 days.",
        "(b) If we were to take many random samples of 1,151 adults and compute a confidence interval for each sample, about 95% of those intervals would contain the true population mean number of mentally unhealthy days.",
        "(c) The 99% confidence interval would be wider than the 95% interval because a higher confidence level requires capturing more of the sampling distribution. The formula is: CI = x̄ ± z* × SE where z* is larger for 99% than for 95%.",
        "(d) The standard error would be larger for a smaller sample size (500 vs 1,151). The formula: SE = s / √n indicates that decreasing n increases SE, all else equal."
      ],
      "conclusion": "Increasing confidence level widens the interval, while reducing sample size increases standard error, reducing precision.",
      "keywords": [
        "confidence interval",
        "mean",
        "standard error",
        "sample size",
      "confidence level"
    ],
    "source": "OpenIntro Statistics, 4th Edition (2019), Example 5.12, Chapter 5, Page 188"
  },

 {
    "index": "C2.345",
    "topic": "Confidence Interval",
    "difficulty": "intermediate",
    "problem": "A manager is about to oversee the mass production of a new tire model in her
factory, and she would like to estimate what proportion of these tires will be rejected through
quality control. The quality control team has monitored the last three tire models produced by
the factory, failing 1.7% of tires in the first model, 6.2% of the second model, and 1.3% of the
third model. The manager would like to examine enough tires to estimate the failure rate of the
new tire model to within about 1% with a 90% confidence level. Perform the sample size
computation for each past model separately.",
    "solution_steps": [
      "Formula: n = (Z^2 * p * (1 - p)) / E^2, where Z = 1.645 for 90% confidence, E = 0.01",
      "Model 1 (p = 0.017): n = (1.645^2 * 0.017 * (1 - 0.017)) / 0.0001 = (2.706 * 0.017 * 0.983) /
0.0001 ≈ 453 tires",
      "Model 2 (p = 0.062): n = (1.645^2 * 0.062 * (1 - 0.062)) / 0.0001 = (2.706 * 0.062 * 0.938) /
0.0001 ≈ 1573 tires",
      "Model 3 (p = 0.013): n = (1.645^2 * 0.013 * (1 - 0.013)) / 0.0001 = (2.706 * 0.013 * 0.987) /
0.0001 ≈ 347 tires"
    ],
    "conclusion": "The sample sizes needed to estimate the failure rate within ±1% at 90%
confidence level are: Model 1 – 453 tires, Model 2 – 1573 tires, Model 3 – 347 tires.",
    "keywords": [
      "sample size calculation",
      "proportion estimation",
      "quality control",
      "margin of error"
    ],
    "source": "OpenIntro Statistics, Fourth Edition (2019), Section 6.1: Inference for a Single
Proportion, Page 213"
  },
{
    "index": "C3.567",
    "topic": "Confidence Intervals",
    "difficulty": "advanced",
    "problem": "You are given a dataset assumed to be i.i.d. from a normal distribution: X₁, ..., X□ ~ N(μ, σ²). Let τ be the 95th percentile of the distribution, i.e., P(X < τ) = 0.95. Suppose the observed data are: 3.23, -2.50, 1.88, -0.68, 4.43, 0.17, 1.03, -0.07, -0.01, 0.76, 1.76, 3.18, 0.33, -0.31, 0.30, -0.61, 1.52, 5.43, 1.54, 2.28, 0.42, 2.33, -1.03, 4.00, 0.39. Tasks: 1. Find the MLE of τ (the 95th percentile). 2. Find the standard error of the MLE using the delta method. 3. Estimate the standard error using the parametric bootstrap.",
    "solution_steps": [
      "Step 1: Let τ = μ + z_{0.95} * σ, where z_{0.95} ≈ 1.645.",
      "Step 2: Estimate μ and σ using MLEs: μ̂ = sample mean ≈ 1.217, σ̂ = sample standard deviation (MLE) ≈ 1.857",
      "Step 3: Compute τ̂ = μ̂ + 1.645 * σ̂ ≈ 1.217 + 1.645 * 1.857 ≈ 4.27",
      "Step 4: Use the delta method to estimate standard error: Var(τ̂) ≈ Var(μ̂) + (1.645)^2 * Var(σ̂), Var(μ̂) = σ² / n, Var(σ̂) ≈ σ² / (2n), SE ≈ σ̂ * sqrt((1 + 1.645² / 2) / n) ≈ 1.857 * sqrt(0.0905) ≈ 0.559",
      "Step 5: Parametric Bootstrap: Generate 1000 samples from N(μ̂, σ̂²), For each, compute τ̂_b = μ̂_b + 1.645 * σ̂_b, Compute standard deviation of τ̂_b values as SE ≈ 0.562"
    ],
    "conclusion": "The MLE of the 95th percentile τ is approximately 4.27. The standard error using the delta method is approximately 0.559, and using the parametric bootstrap is approximately 0.562.",
    "keywords": [
      "MLE",
      "percentile",
      "normal distribution",
      "delta method",
      "parametric bootstrap",
      "confidence interval"
    ],
    "source": "All of Statistics by Larry Wasserman, Chapter 9: Parametric Inference, Page 146"
  },

  {
    "index": "D1.234",
    "topic": "Sampling Distribution and Central Limit Theorem",
    "difficulty": "basic",
    "problem": "Suppose that the height of men has mean 68 inches and standard deviation 2.6 inches. We draw a random sample of 100 men. Find (approximately) the probability that the average height of men in our sample will be at least 68 inches.",
    "solution_steps": [
      "Step 1: Identify population parameters: Mean (μ) = 68 inches, Standard deviation (σ) = 2.6 inches, Sample size (n) = 100",
      "Step 2: Use the Central Limit Theorem: The sampling distribution of the sample mean X̄ is approximately normal with: Mean = μ = 68, Standard error = σ / √n = 2.6 / √100 = 0.26",
      "Step 3: Standardize the value 68 to a Z-score: Z = (68 - 68) / 0.26 = 0",
      "Step 4: Compute the probability: P(X̄ ≥ 68) = P(Z ≥ 0) = 0.5"
    ],
    "conclusion": "There is approximately a 50% probability that the average height of the 100 randomly selected men will be at least 68 inches.",
    "keywords": [
      "sampling distribution",
      "central limit theorem",
      "standard error",
      "normal approximation",
      "probability"
    ],
    "source": "All of Statistics by Larry Wasserman, Chapter 5.8: Exercises, Exercise 83, Page 83"
  },
   {
{
  "index": "D2.345",
  "topic": "Parameter Estimation",
  "difficulty": "intermediate",
  "problem": "A multiple choice exam has 4 choices per question. A student has a 0.5 probability of knowing the answer, a 0.25 probability of being able to eliminate one wrong choice, and a 0.25 probability of guessing blindly. If they know the answer, they will get it correct. If not, they guess from 3 or 4 options. If a student answers a question correctly, what is the probability they actually knew the answer?",
  "solution_steps": [
    "Step 1: Define events – C: student answers correctly, K: student knows the answer, E: student can eliminate one choice, G: student guesses blindly.",
    "Step 2: Use Bayes’ Theorem to compute P(K | C):",
    "P(K | C) = [P(C | K) * P(K)] / [P(C | K) * P(K) + P(C | E) * P(E) + P(C | G) * P(G)]",
    "Step 3: Plug in values:",
    "P(C | K) = 1 (always correct if they know the answer), P(K) = 0.5",
    "P(C | E) = 1/3 (guess from 3 options), P(E) = 0.25",
    "P(C | G) = 1/4 (guess from 4 options), P(G) = 0.25",
    "Step 4: Compute numerator: 1 * 0.5 = 0.5",
    "Step 5: Compute denominator: 0.5 + (1/3 * 0.25) + (1/4 * 0.25) = 0.5 + 0.0833 + 0.0625 = 0.6458",
    "Step 6: Final result: P(K | C) = 0.5 / 0.6458 ≈ 0.774"
  ],
  "conclusion": "If a student answers correctly, there is approximately a 77.4% chance they actually knew the answer.",
  "keywords": [
    "Bayes' Theorem",
    "conditional probability",
    "multiple choice exam",
    "parameter estimation"
  ],
```

  "source": "Example 4.9, Chapter 4 – Probability, University of Connecticut OER: https://probability.oer.math.uconn.edu/wp-content/uploads/sites/2187/2018/01/prob3160ch4.pdf"
},

  {
  "index": "D2.456",
  "topic": "Sampling Distributions",
  "difficulty": "basic",
  "problem": "The mean and standard deviation of the tax value of all vehicles registered in a certain state are μ = $13,525 and σ = $4,180. Suppose random samples of size 100 are drawn from this population. What are the mean μ̄X and standard deviation σ̄X of the sample mean X̄?",
  "solution_steps": [
    "Step 1: Identify the population parameters: μ = 13,525 and σ = 4,180.",
    "Step 2: Use the formula for the mean of the sampling distribution of the sample mean: μ̄X = μ = 13,525.",
    "Step 3: Use the formula for the standard deviation of the sample mean: σ̄X = σ / √n.",
    "Step 4: Plug in the values: σ̄X = 4,180 / √100.",
    "Step 5: Compute: √100 = 10, so σ̄X = 4,180 / 10 = 418."
  ],
  "conclusion": "The mean of the sample mean is $13,525 and the standard deviation of the sample mean is $418.",
  "keywords": [
    "sampling distribution",
    "mean of sample mean",
    "standard deviation",
    "central limit theorem"
  ],
  "source": "Saylor Academy, 'Introductory Statistics – Sampling Distributions': https://saylordotorg.github.io/text_introductory-statistics/s10-sampling-distributions.html"
},

```
{
  "index": "D2.567",
  "topic": "Confidence Intervals",
  "difficulty": "intermediate",
  "problem": "There were 2430 Major League Baseball (MLB) games played in 2009, and the home team won in 54.9% of the games. If we consider the games played in 2009 as a sample of all MLB games, find and interpret a 90% confidence interval for the proportion of games the home team wins in Major League Baseball.",
  "solution_steps": [
    "Step 1: Identify sample proportion p̂ = 0.549 and sample size n = 2430.",
    "Step 2: Use z* = 1.645 for a 90% confidence level.",
    "Step 3: Use the confidence interval formula: p̂ ± z* · sqrt[p̂(1 - p̂)/n].",
    "Step 4: Compute the standard error: sqrt[0.549 × 0.451 / 2430] ≈ 0.0103.",
    "Step 5: Multiply by z*: 1.645 × 0.0103 ≈ 0.017.",
    "Step 6: Construct the interval: 0.549 ± 0.017 → (0.532, 0.566)."
  ],
  "conclusion": "We are 90% confident that the proportion of MLB games that are won by the home team is between 0.532 and 0.566. This assumes the 2009 season is representative of all MLB games.",
  "keywords": [
    "confidence interval",
    "sample proportion",
    "standard error",
    "z-score",
    "baseball statistics"
  ],
  "source": "Based on: 'STAT 302 Homework Solutions', University of Wisconsin: https://pages.stat.wisc.edu/~larget/stat302/sol08.pdf"
}
```

{

"index": "E1.123",

"topic": "Hypothesis Testing",

"difficulty": "basic",

"problem": "For each of the following assertions, state whether it is a legitimate statistical hypothesis and why: a. H_0: $\sigma = 100$ b. H_0: $\sigma \geq 45$ c. H_0: $\sigma \leq 20$ d. H_0: $\sigma_1 / \sigma_2 = 1$ e. H_0: $X^2 \leq 55$ f. H_0: $\lambda \neq 0.01$, where λ is the parameter of an exponential distribution used to model component lifetime.",

"solution_steps": [

"1. For each assertion, check if it follows the form of a statistical hypothesis (H_0: parameter = value or H_0: parameter $\leq$ value or H_0: parameter $\geq$ value).",

"2. Determine if the hypothesis involves parameters and values appropriate for statistical testing.",

"3. Validate if the assertions are correctly stated for testing purposes.",
"a. H_0: $\sigma = 100$ - LEGITIMATE: Tests a specific value for the population standard deviation",
"b. H_0: $\sigma \geq 45$ - LEGITIMATE: Tests if standard deviation is at least 45",
"c. H_0: $\sigma \leq 20$ - LEGITIMATE: Tests if standard deviation is at most 20",
"d. H_0: $\sigma_1 / \sigma_2 = 1$ - LEGITIMATE: Tests equality of two population variances",
"e. H_0: $X^2 \leq 55$ - NOT LEGITIMATE: X^2 refers to a random variable, not a parameter",
"f. H_0: $\lambda \neq 0.01$ - NOT LEGITIMATE: Null hypothesis should use equality, this should be the alternative hypothesis"
],
"conclusion": "Assertions a, b, c, and d are legitimate statistical hypotheses because they involve population parameters. Assertions e and f are not legitimate as written.",
"keywords": [
"hypothesis testing",
"null hypothesis",
"population parameters",
"statistical inference"
],
"source": "Probability and Statistics for Engineering and the Sciences, 9th Edition by Jay L. Devore, Chapter 8, Question 1E, page 325 (https://www.vaia.com/en-us/textbooks/math/probability-and-statistics-for-engineering-and-sciences-9th/tests-of-hypotheses-based-on-a-single-sample/q1e-for-each-of-the-following-assertions-state-whether-it-is/)"
},

{
"Index": “E1.234”,
"topic": "Hypothesis Testing",
"difficulty": "intermediate",
"problem": "A motor manufacturer wishes to replace steel suspension components with aluminium ones to improve performance and fuel efficiency. Tensile strength tests are carried out on samples from two components. The data is as follows:\n\nComponent 1: n = 15, mean = 90 kg/mm^2, standard deviation = 2.3\nComponent 2: n = 10, mean = 88 kg/mm^2, standard deviation = 2.2\n\nAt the 5% significance level, is there a statistically significant difference in the tensile strengths of the two components?",
"solution_steps": [
"Step 1: Define null and alternative hypotheses:\nH0: $\mu 1 - \mu 2 = 0$ (no difference)\nH1: $\mu 1 - \mu 2 \neq 0$ (difference exists)",
"Step 2: Use the formula for the two-sample Z-test statistic:\nZ = $(\bar{X}1 - \bar{X}2 - (\mu 1 - \mu 2))$ / sqrt($\sigma 1^2$/n1 + $\sigma 2^2$/n2)",

```
    "Step 3: Plug in values:\nZ = (90 − 88 − 0) / sqrt((2.3)²/15 + (2.2)²/10)\nZ = 2 / sqrt(0.3527 + 0.484) ≈ 2 / 0.918 = 2.18",
    "Step 4: Compare with critical value:\nAt 5% significance level (two-tailed), z* ≈ 1.96\nSince 2.18 > 1.96, we reject the null hypothesis"
  ],
  "conclusion": "There is sufficient evidence at the 5% significance level to suggest a difference in tensile strength between the two components. However, further tests are recommended before making a final decision.",
  "keywords": [
    "two-sample hypothesis test",
    "tensile strength",
    "z-test",
    "level of significance",
    "standard deviation"
  ],
"source": "University of Sheffield – HELM Workbook, Section 41.3: 'Tests Concerning Two Samples': https://www.sheffield.ac.uk/media/32113/download?attachment"
},




     {
        "index": "E2.345",
        "topic": "Hypothesis Testing",
        "difficulty": "intermediate",
        "problem": "In a psychology class, the average grade is known to be 67.5 points with a population standard deviation of 9.5. A researcher samples 20 psychology students and finds their average grade is 72.3. Test at the 5% significance level whether psychology students score differently than the class average.",
        "solution_steps": [
          "1. Define hypotheses: H₀: μ = 67.5, H₁: μ ≠ 67.5.",
          "2. Given: sample mean x̄ = 72.3, σ = 9.5, n = 20, α = 0.05.",
          "3. Compute standard error: SE = σ / √n = 9.5 / √20 ≈ 2.12.",
          "4. Compute z-score: z = (x̄ − μ₀) / SE = (72.3 − 67.5) / 2.12 ≈ 2.26.",
          "5. For a two-tailed test at α = 0.05, critical values are ±1.96.",
          "6. Since 2.26 > 1.96, reject H₀.",
          "7. Conclusion: There is sufficient evidence to conclude that psychology students’ grades differ significantly from the class average."
        ],
        "conclusion": "The test indicates a significant difference (at 5% level) between psychology students and the class average.",
  "keywords": [
    "one-sample z-test",
    "hypothesis testing",
    "psychology students",
    "grade comparison"
  ],
  "source": "Adapted from LibreTexts example ‘One Sample Mean z Test’, Stats 200 ()"
},
```

{
  "index": "E2.456",
  "topic": "ANOVA",
  "difficulty": "intermediate",
  "problem": "A nutritionist tests the effects of three different meal plans (Plan A, Plan B, Plan C) on weight loss over a 6-week period. The weight loss (in pounds) for each plan is as follows: Plan A = [6, 7, 8], Plan B = [5, 6, 7], Plan C = [8, 9, 10]. Conduct a one-way ANOVA test.",
  "solution_steps": [
    "Step 1: Calculate the group means: Mean of Plan A = (6 + 7 + 8) / 3 = 7, Mean of Plan B = (5 + 6 + 7) / 3 = 6, Mean of Plan C = (8 + 9 + 10) / 3 = 9",
    "Step 2: Calculate the overall mean: Overall mean = (6 + 7 + 8 + 5 + 6 + 7 + 8 + 9 + 10) / 9 = 66 / 9 = 7.3333",
    "Step 3: Calculate the sum of squares between groups (SSB): SSB = n × Σ(group_mean - overall_mean)^2. SSB = 3 × [(7 - 7.3333)^2 + (6 - 7.3333)^2 + (9 - 7.3333)^2] = 3 × [0.1111 + 1.7778 + 2.7778] = 3 × 4.6667 = 14",
    "Step 4: Calculate the sum of squares within groups (SSW): SSW = Σ(individual_value - group_mean)^2. Plan A: (6-7)^2 + (7-7)^2 + (8-7)^2 = 1 + 0 + 1 = 2. Plan B: (5-6)^2 + (6-6)^2 + (7-6)^2 = 1 + 0 + 1 = 2. Plan C: (8-9)^2 + (9-9)^2 + (10-9)^2 = 1 + 0 + 1 = 2. SSW = 2 + 2 + 2 = 6",
    "Step 5: Calculate the degrees of freedom: df_between = k - 1 = 3 - 1 = 2, df_within = N - k = 9 - 3 = 6",
    "Step 6: Calculate the mean squares: MSB = SSB / df_between = 14 / 2 = 7, MSW = SSW / df_within = 6 / 6 = 1",
    "Step 7: Calculate the F-statistic: F = MSB / MSW = 7 / 1 = 7",
    "Step 8: Determine the critical value: At α = 0.05 with df_1 = 2 and df_2 = 6, F_critical = 5.143",
    "Step 9: Make decision: Since F = 7 > F_critical = 5.143, we reject the null hypothesis H_0 that all group means are equal"
  ],
  "conclusion": "The F-statistic of 7 exceeds the critical value of 5.143 (α = 0.05), so we reject the null hypothesis. There are statistically significant differences in weight loss among the three meal plans. Post-hoc tests would be needed to determine which specific plans differ from each other.",
  "keywords": [
    "ANOVA",
    "meal plans",
    "weight loss analysis",
    "nutritional study",
    "F-test",
    "hypothesis testing"
  ],
  "source": "Daniel Giurleo (#0788070), SOCI-3160 Week 8 Exercises – Draft 01, Trent University, Course: SOCI 3160H, Nov 9, 2024, Page 4 (https://www.cliffsnotes.com/study-notes/21973801)"
}

```
  {
    "index": "E2.567",
    "topic": "Hypothesis Testing (Proportions)",
    "difficulty": "intermediate",
    "problem": "A newspaper reported that 75% of students regularly cycle to college. A college dean believes that figure to be different at his college. He asks a sample of 160 students, and 109 say they do cycle. Test the dean's belief at the 5% level of significance.",
    "solution_steps": [
      "1. State the hypotheses: H₀: p = 0.75 (the proportion is 75%), H₁: p ≠ 0.75 (the proportion is different)",
      "2. Model the distribution: X ~ B(160, 0.75) → approximated by normal: N(μ = 120, σ = √(160 × 0.75 × 0.25) = √30 ≈ 5.48)",
      "3. Apply continuity correction: Use 109.5 instead of 109.",
      "4. Calculate Z-score: Z = (109.5 − 120) / √30 = −1.917",
      "5. Find probability: P(X ≤ 109) = P(Z ≤ −1.917) = 0.0277",
      "6. Compare with significance level: Since this is a two-tailed test, compare with 0.025 (half of 5%). 0.0277 > 0.025 → do not reject H₀"
    ],
    "conclusion": "Accept H₀. There is insufficient evidence to show that the proportion of students cycling at the college is different from the national average of 75%.",
    "keywords": [
      "hypothesis test",
      "proportion",
      "normal approximation",
      "continuity correction",
      "two-tailed test"
    ],
    "source": "Cambridge International AS & A Level Mathematics: Probability & Statistics 2, Worked Example 1.2"
  },
  {
    "index": "E2.678",
    "topic": "Hypothesis Testing - Binomial to Normal Approximation",
    "difficulty": "intermediate",
    "problem": "In a test, the questions are all multiple choice with five possible options. (a) In a test with twelve questions, one student gets four questions correct. Test, at the 10% significance level, the null hypothesis that the student is guessing the answers. (b) In a further test there are 120 questions. The same student took the test and got 32 correct. Test, at the 10% significance level, whether there is evidence to show the student is guessing the answers.",
    "solution_steps": [
      "(a) Let X be the number of correct answers achieved.",
       "X ~ B(12, 0.2)",
       "H₀: p = 0.2, H₁: p > 0.2",
       "P(X ≥ 4) = 1 – P(X ≤ 3)",
       "P(X ≤ 3) = P(X=0) + P(X=1) + P(X=2) + P(X=3)",
       "Using binomial terms: = 1 – [ (12C0)(0.2^0)(0.8^12) + (12C1)(0.2^1)(0.8^11) + (12C2)(0.2^2)(0.8^10) + (12C3)(0.2^3)(0.8^9) ] = 1 – 0.795 = 0.205",
       "0.205 > 0.1. Accept H₀. There is sufficient evidence to show that the student is guessing answers."

     "(b) X ~ B(120, 0.2) ≈ N(24, 19.2)",
       "H₀: p = 0.2 or μ = 24, H₁: p > 0.2 or μ > 24",
       "Use normal approximation with continuity correction: P(X ≥ 32) = P(X ≥ 31.5)",
       "Standardize: Z = (31.5 – 24) / √19.2 ≈ 1.712",
       "P(Z ≥ 1.712) = 1 – Φ(1.712) = 1 – 0.957 = 0.043",
       "0.043 < 0.1. Reject H₀. There is evidence to suggest that the student might not be guessing answers."
    ],
    "conclusion": "For the 12-question test, we accept that the student is guessing. For the 120-question test, we reject the guessing hypothesis.",
    "keywords": [
     "binomial distribution",
     "normal approximation",
     "significance test",
     "continuity correction",
     "hypothesis testing",
     "p-value"
    ],
    "source": "Cambridge International AS & A Level Mathematics: Probability & Statistics 2, Worked Example 1.1"
  },
  {
    "index": "E2.789",
    "topic": "Hypothesis testing",
    "difficulty": "intermediate",
    "problem": "Georgianna claims that in a small city renowned for its music school, the average child takes less than 5 years of piano lessons. A random sample of 24 children from the city shows a mean of 4.6 years of piano lessons with a standard deviation of 2.2 years. Evaluate Georgianna's claim using a hypothesis test at a 0.05 significance level.",
    "solution_steps": [
      "Step 1: Set up hypotheses: H₀: μ ≥ 5 (null hypothesis - average is 5 years or more), H₁: μ < 5 (alternative hypothesis - Georgianna's claim)",
      "Step 2: Given values: sample_size = 24, sample_mean = 4.6, sample_sd = 2.2, alpha = 0.05, test_type = left-tailed",
      "Step 3: Compute test statistic (t): t = (x̄ - μ₀) / (s / √n) = (4.6 - 5) / (2.2 / √24) = -0.4 / 0.449 ≈ -0.891",
      "Step 4: Determine critical value and p-value: degrees_of_freedom = 23, critical_value = -1.714, p_value ≈ 0.191",
      "Step 5: Make decision: Since t = -0.891 > -1.714 and p-value = 0.191 > 0.05, we fail to reject H₀."
    ],
    "conclusion": "There is insufficient evidence to support Georgianna's claim that the average child takes less than 5 years of piano lessons.",
    "keywords": [
      "t-test",
      "hypothesis testing",
      "small sample",
      "mean comparison",
      "one-sample test"
    ],
    "source": "OpenIntro Statistics, 4th Edition (2019), Section 7.1: One-Sample Means with the t-Distribution, Page 261"
  },
  {
    "index": "E2.890",
    "topic": "One-Sample Z-Test (Known σ)",
    "difficulty": "intermediate",
    "problem": "Jeffrey, an eight-year-old, had an established mean time of 16.43 seconds for swimming the 25-yard freestyle, with a population standard deviation of 0.8 seconds. His father Frank bought him expensive goggles, suspecting they might help him swim faster. In 15 timed swims using the goggles, Jeffrey's sample mean time was 16 seconds. Test whether the goggles helped Jeffrey swim faster, using a significance level of α = 0.05. Assume the swim times are normally distributed.",
    "solution_steps": [
      "Step 1: State the hypotheses. Null Hypothesis (H₀): μ = 16.43, Alternative Hypothesis (Hₐ): μ < 16.43. This is a left-tailed test since we're checking if Jeffrey swims faster (i.e., mean time is less).",
      "Step 2: Define the distribution. Random variable: X̄ = mean swim time. Since σ is known and the data are normal, use the z-distribution. Parameters: μ = 16.43, σ = 0.8, n = 15",
      "Step 3: Calculate the test statistic. Standard Error (SE) = σ / √n = 0.8 / √15 ≈ 0.206. z = (16 - 16.43) / 0.206 ≈ -2.087",
      "Step 4: Find the p-value. Using the standard normal distribution: p-value = P(Z < -2.087) ≈ 0.0187",
      "Step 5: Compare p-value and α. α = 0.05. Since p-value = 0.0187 < 0.05, we reject the null hypothesis.",
      "Step 6: Conclusion. There is sufficient evidence at the 5% significance level to conclude that Jeffrey's mean time is less than 16.43 seconds. The goggles likely helped him swim faster."
    ],
    "conclusion": "Reject H₀. The data provide sufficient evidence to support that Jeffrey swims the 25-yard freestyle faster (i.e., in less than 16.43 seconds) using the new goggles.",
    "interpretation": "A p-value of 0.0187 means that if Jeffrey's true average time were still 16.43 seconds, there is only a 1.87% chance of observing a sample mean of 16 seconds or lower. This is unlikely under H₀, hence we reject it.",
 "errors": {
   "type_I_error": "Concluding Jeffrey swims faster (μ < 16.43) when he actually doesn't (μ = 16.43).",
   "type_II_error": "Failing to conclude Jeffrey swims faster when he actually does (μ < 16.43)."
  },
    "keywords": [
      "hypothesis testing",
      "z-test",
      "left-tailed test",
      "p-value",
      "normal distribution",
      "swimming performance"
    ],
    "source": "OpenStax: Introductory Statistics 2e, Section 9.5"
  },
  {
    "index": "E3.234",
    "topic": "One Sample t-test",
    "difficulty": "advanced",
    "problem": "Use the data [0.86, 1.53, 1.57, 1.81, 0.99, 1.09, 1.29, 1.78, 1.29, 1.58], comprising a sample of n = 10 lactic acid measurements in cheese. The lactic acid measurements are a random sample from a normal distribution with unknown mean μ and unknown variance σ². Hypotheses: H₀: μ ≤ 1.2, H₁: μ > 1.2. Tasks: a. Perform the level α₀ = 0.05 test of these hypotheses. b. Compute the p-value.",
    "solution_steps": [
      "Compute sample size: n = 10",
      "Compute sample mean: x̄ = 1.379",
      "Compute sample standard deviation: s ≈ 0.3277",
      "Calculate test statistic: t = (1.379 - 1.2) / (0.3277 / sqrt(10)) ≈ 1.831",
      "Degrees of freedom: df = 9",
      "Find critical value for one-tailed test at α = 0.05: t_critical ≈ 1.833",
      "Compare t = 1.831 < t_critical = 1.833 → Fail to reject H₀",
      "p-value ≈ 0.049 < p < 0.05 (from t-distribution table with df = 9)"
    ],
    "conclusion": "We fail to reject the null hypothesis at the 0.05 level. There is insufficient evidence to conclude that the mean lactic acid concentration exceeds 1.2.",
    "keywords": [
      "t-test",
      "hypothesis testing",
      "one-sample test",
      "lactic acid",
      "small sample",
      "cheese analysis"
    ],
    "source": "Probability and Statistics (Morris H. DeGroot, Mark J. Schervish), Chapter 8, Example 8.5.4, Page 512"
  },
```

{
"index": "E3.345",
"topic": "Hypothesis Testing and Nonparametric Inference",
"difficulty": "advanced",
"problem": "In 1861, 10 essays appeared in the New Orleans Daily Crescent signed 'Quintus Curtius Snodgrass.' Some suspected they were actually written by Mark Twain. To investigate, we consider the proportion of three-letter words in the texts. From 8 Twain essays: [0.225, 0.262, 0.217, 0.240, 0.230, 0.229, 0.235, 0.217]. From 10 Snodgrass essays: [0.209, 0.205, 0.196, 0.210, 0.202, 0.207, 0.224, 0.223, 0.220, 0.201]. (a) Perform a Wald test for equality of the means. Use the nonparametric plug-in estimator (difference of the sample means). Report the p-value and a 95% confidence interval for the difference of means. What do you conclude? (b) Use a permutation test to avoid the large-sample approximation. What is your conclusion?",
"solution_steps": [
"(a) Wald Test:
"Calculate sample means: Twain (x̄1) = 0.231875, Snodgrass (x̄2) = 0.2097",
"Calculate sample variances: s1^2 = 0.0002846, s2^2 = 0.0000892",
"Compute standard error (SE): SE = sqrt(s1^2/n1 + s2^2/n2) = sqrt(0.0002846/8 + 0.0000892/10) ≈ 0.006670",
"Compute test statistic: z = (x̄1 - x̄2)/SE = (0.231875 - 0.2097)/0.006670 ≈ 3.324",
"Compute p-value: p ≈ 0.0008 (two-tailed test)",
"95% CI = (x̄1 - x̄2) ± z* × SE = 0.022175 ± 1.96 × 0.006670 ≈ [0.0091, 0.0352]

Since the p-value is very small (< 0.01), we reject H0. The 95% CI does not contain 0. There is strong evidence that the mean proportion of three-letter words differs between the two authors.",

"(b) Permutation Test:

"Combine all 18 proportions into one pool.",
"Randomly permute the combined set and split into two groups (n1=8, n2=10).",
"Compute the difference in means for each permutation.",
"Repeat this process 10,000 times to generate the permutation distribution.",
"Count how often |permuted difference| ≥ |observed difference| (|0.231875 - 0.2097| = 0.022175).",
"Estimated p-value = (number of |permuted diffs| ≥ 0.022175) / 10,000

The permutation test also yields a very small p-value (typically ≈ 0.0008), confirming the Wald test result without relying on normality assumptions. There is strong evidence that the essays differ in their writing style, casting doubt on the Twain authorship of Snodgrass."
],
"conclusion": "Both the Wald test and permutation test indicate a statistically significant difference in the proportion of three-letter words. This supports the hypothesis that the Snodgrass essays are not written by Twain.",
"keywords": [

```
    "wald test",
    "permutation test",
    "confidence interval",
    "difference of means",
    "nonparametric inference",
    "text analysis",
    "authorship attribution"
  ],
  "source": "Adapted from Wasserman, All of Statistics, Section 10.11, p. 171"
},
  {
    "index": "E3.456",
    "topic": "Hypothesis Testing – Binomial Distribution",
    "difficulty": "advanced",
    "problem": "A paper company finds that, over time, defects occur at a rate of 1 in every 250 sheets of paper produced (p = 0.004). After switching to recycled paper, the company examines a sample of 300 sheets and finds 5 defective sheets. An employee conducts a hypothesis test at the 5% level of significance to assess whether the switch has affected the defect rate. The employee assumes X ~ B(300, p), sets H₀: p = 0.004 and H₁: p ≠ 0.004, and calculates P(X = 5 | p = 0.004) = 0.9985. Based on this, they conclude: 'Since 0.9985 > 0.05, accept H₀. There is no evidence that the defect rate has changed.' Explain the mistakes and determine the correct conclusion.",
    "solution_steps": [
      "Step 1: Identify hypotheses: H₀: p = 0.004, H₁: p ≠ 0.004",
      "Step 2: Note that X ~ B(300, 0.004), with expected value μ = 300 × 0.004 = 1.2 and standard deviation σ = sqrt(300 × 0.004 × 0.996) ≈ 1.0933",
      "Step 3: Employee's first mistake: Using P(X = 5) instead of computing a two-tailed p-value. A hypothesis test requires calculating P(X ≥ 5 or X ≤ x_lower) or a z-score-based approach.",
      "Step 4: Employee's second mistake: Interpreting P(X = 5) > 0.05 to 'accept H₀'. This is incorrect logic and language; we never 'accept' H₀ — we 'fail to reject' or 'reject' it based on the p-value.",
      "Step 5: Apply normal approximation with continuity correction. Use Z = (4.5 - 1.2) / 1.0933 ≈ 3.02",
      "Step 6: Compute two-tailed p-value: P(|Z| > 3.02) ≈ 2 × 0.00126 = 0.00252",
      "Step 7: Since p-value ≈ 0.0025 < 0.05, reject H₀"
    ],
    "conclusion": "There is significant evidence at the 5% level that the defect rate has changed after switching to recycled paper. The employee incorrectly used a point probability and misinterpreted the test result.",
    "keywords": [
      "hypothesis testing",
      "binomial distribution",
      "p-value interpretation",
      "normal approximation",
      "defect rate analysis"
    ],
    "source": "Edexcel A-Level Statistics and Mechanics 1 Practice Book, Hypothesis Testing Question (via https://james28.uk/School/Active%20Learn/Books/Single/Practice/Stats%20and%20Mechanics%201%20Practice.pdf)"
  },
 {
  "index": "E3.567",
  "topic": "One-Way ANOVA – Hospital ICU Hours",
  "difficulty": "advanced",
  "keywords": [
    "ANOVA",
    "F-test",
    "hospital comparison",
    "ICU hours",
    "healthcare statistics"
  ],
  "problem": "The Eastside Health Authority collects data on hours spent in intensive care by
patients with suspected coronary heart attacks across five hospitals (A–E). The data are as
follows: Hospital A: 30, 25, 12, 23, 16; Hospital B: 42, 57, 47, 30; Hospital C: 65, 46, 55, 27;
Hospital D: 67, 58, 81; Hospital E: 70, 63, 80. Use a one-factor analysis of variance to test, at
the 1% level of significance, whether there is a difference in mean ICU hours between
hospitals.",
  "correct_solution": {
    "steps": [
      "Step 1: Organize data by group and compute group totals and means.",
      "Step 2: Compute grand total T = 894 and grand mean = 894 divided by 19, approximately
47.05.",
      "Step 3: Compute total sum of squares (SST): sum of squares of all observations = 50354;
then SST = 50354 minus square of 894 divided by 19, which is approximately 8288.95.",
      "Step 4: Compute between-group sum of squares (SSB): sum of squares of each group total
divided by group size, minus square of 894 divided by 19, which is approximately 6506.73.",
      "Step 5: Compute within-group sum of squares (SSW): SSW = SST minus SSB = 8288.95
minus 6506.73 = 1782.22.",
      "Step 6: Compute degrees of freedom: degrees of freedom between groups = 4, degrees of
freedom within groups = 14.",
      "Step 7: Compute mean squares: MSB = 6506.73 divided by 4 = 1626.68; MSW = 1782.22
divided by 14, approximately 127.30.",
      "Step 8: Compute F-statistic: F = MSB divided by MSW = 1626.68 divided by 127.30,
approximately 12.78.",
      "Step 9: Compare with critical value: F critical at 1% significance for (4, 14) degrees of
freedom is approximately 4.60.",
      "Step 10: Since F > F_critical, reject the null hypothesis."
    ],
    "conclusion": "At the 1% significance level, there is strong evidence that the mean ICU hours
differ between the five hospitals."
  },
  "source": "MEI/CIMT A-Level Further Statistics 1, Chapter 7: Analysis of Variance, Exercise
7A, Question 5"
}
```

```
  {
    "index": "F1.234",
    "topic": "Bayes' Theorem",
    "difficulty": "basic",
    "problem": "A spam filter is designed by looking at commonly occurring phrases in spam. Suppose that 80% of email is spam. In 10% of the spam emails, the phrase 'free money' is used, whereas this phrase is only used in 1% of non-spam emails. A new email has just arrived, which does mention 'free money'. What is the probability that it is spam?",
    "solution_steps": [
      "Let S be the event that an email is spam.",
      "Let F be the event that an email has the 'free money' phrase.",
      "We are required to compute P(S | F).",
      "Apply Bayes' Rule:",
      "P(S | F) = [P(F | S) * P(S)] / P(F)",
      "P(F | S) = 0.1, P(S) = 0.8, P(F | not S) = 0.01, P(not S) = 0.2",
      "P(F) = (0.1 * 0.8) + (0.01 * 0.2) = 0.08 + 0.002 = 0.082",
      "P(S | F) = (0.1 * 0.8) / 0.082 = 0.08 / 0.082 = 80 / 82 ≈ 0.9756"
    ],
    "conclusion": "The probability that an email is spam given it contains the phrase 'free money' is approximately 0.9756.",
    "keywords": [
      "Bayes' Theorem",
      "spam filtering",
      "conditional probability",
      "phrase detection",
      "posterior probability"
    ],
    "source": "Introduction to Probability by Joseph K. Blitzstein and Jessica Hwang, Chapter 2, Page 11"
   },
  {
    "index": "F2.345",
    "topic": "Bayesian Statistics",
    "difficulty": "intermediate",
    "problem": "Given a Beta(5, 3) prior and observing 25 successes and 10 failures, find the posterior distribution parameters and compute the posterior variance.",
    "solution_steps": [
      "Step 1: Prior distribution is Beta(5, 3).",
      "Step 2: Observed data: 25 successes, 10 failures.",
      "Step 3: Update the prior with observed data: Posterior distribution = Beta(5 + 25, 3 + 10) = Beta(30, 13).",
      "Step 4: Compute the posterior mean: Mean = α / (α + β) = 30 / (30 + 13) = 30 / 43 ≈ 0.698.",
      "Step 5: Compute the posterior variance: Variance = (α * β) / [(α + β)^2 * (α + β + 1)] = (30 * 13) / [43^2 * 44] = 390 / 81,004 ≈ 0.004815.",
      "Step 6: Thus, the posterior distribution is Beta(30, 13) and the posterior mean is approximately 0.698."
    ],
    "conclusion": "The posterior distribution is Beta(30, 13) with posterior mean approximately 0.698 and posterior variance approximately 0.004815.",
    "keywords": [
      "Bayesian statistics",
      "Beta distribution",
      "posterior distribution",
      "posterior mean",
      "posterior variance"
    ],
    "source": "MIT OpenCourseWare: 18.05 Introduction to Probability and Statistics, Spring 2022, Lecture 15"
  },
  {
    "index": "F3.456",
    "topic": "Bayes' Theorem – Inverted Probability",
    "difficulty": "advanced",
    "problem": "In Canada, about 0.35% of women over 40 will develop breast cancer in any given year. A common screening test for cancer is the mammogram, but it is not perfect. In 11% of patients with breast cancer, the test gives a false negative result. In 7% of patients without breast cancer, the test gives a false positive result. If we test a random woman over 40 for breast cancer using a mammogram and the test comes back positive, what is the probability that the patient actually has breast cancer?",
    "solution_steps": [
      "1. Define Events: Let BC = patient has breast cancer; NoBC = patient does not have breast cancer; M+ = mammogram result is positive.",
    "2. Assign known probabilities:",
    "- P(BC) = 0.0035 (base rate of breast cancer)",
    "- P(NoBC) = 1 - P(BC) = 0.9965",
    "- P(M+ | BC) = 0.89 (true positive rate)",
    "- P(M+ | NoBC) = 0.07 (false positive rate)",
    "3. Apply Bayes' Theorem to compute the inverted probability:",
    "P(BC | M+) = [P(M+ | BC) * P(BC)] / P(M+)",
    "4. Compute the numerator: P(M+ and BC) = P(M+ | BC) * P(BC) = 0.89 * 0.0035 = 0.00312",
    "5. Compute the denominator using total probability:",
    "P(M+) = P(M+ and BC) + P(M+ and NoBC)",
    "= 0.00312 + (0.07 * 0.9965) = 0.00312 + 0.06976 = 0.07288",
    "6. Compute the final conditional probability:",
    "P(BC | M+) = 0.00312 / 0.07288 ≈ 0.0428"
    ],
    "conclusion": "Even after a positive mammogram, the probability that a woman over 40 actually has breast cancer is only about 4.3%.",
    "explanation": "Although the mammogram is reasonably accurate, the low base rate (prevalence) of breast cancer in the population drastically affects the posterior probability. This is a classic example of the base rate fallacy: people often overestimate the likelihood of disease based on a single positive result, without accounting for how rare the disease is.",
    "keywords": [
     "Bayes' Theorem",
     "conditional probability",
        "false positive",
        "false negative",
        "diagnostic test",
        "base rate fallacy",
        "inverted probability"
      ],
      "source": "OpenIntro Statistics, Fourth Edition by Diez, Barr, Çetinkaya-Rundel, Example 3.42"
    },
 {
  "index": "F3.567",
  "topic": "Bayesian Inference with Rounded Data",
  "difficulty": "advanced",
  "source": "Bayesian Data Analysis, 3rd Edition (Gelman et al., 2013), Exercise 5.9, Chapter 5",
  "problem": "Rounded data: it is a common problem for measurements to be observed in rounded form. Suppose we weigh an object five times and measure weights, rounded to the nearest pound, of 10, 10, 12, 11, 9. Assume the unrounded measurements are normally distributed with a noninformative prior distribution on the mean μ and variance σ².(a) Give the posterior distribution for (μ, σ²) obtained by pretending that the observations are exact unrounded measurements.(b) Give the correct posterior distribution for (μ, σ²) treating the measurements as rounded.(c) How do the incorrect and correct posterior distributions differ? Compare means, variances, and contour plots.",
  "solution": {
        "a": "We treat the observations as exact: y = [10, 10, 12, 11, 9]. With a noninformative prior for (μ, σ²), we use the conjugate normal-inverse-gamma model. Sample mean ȳ = 10.4, sample variance s² = 1.3, and n = 5. Thus, the posterior is:\n- μ | y ~ N(10.4, σ² / 5)\n- σ² | y ~ InvGamma(2, 2.6)",
        "b": "For rounded data, each observation y represents a latent true value z in the interval [y-0.5, y+0.5]. The likelihood becomes:\nL(μ, σ²) = ∏[Φ((yᵢ+0.5−μ)/σ) − Φ((yᵢ−0.5−μ)/σ)]\nThis yields a non-conjugate posterior requiring numerical methods (e.g., MCMC) to sample from. The posterior no longer has a closed form.",
        "c": "Treating rounded data as exact underestimates uncertainty. The mean estimate from the exact model is 10.4. The rounded model typically gives a similar but slightly lower mean, and the posterior variance is larger due to added uncertainty from rounding. The posterior shape for the exact case is elliptical, while for the rounded case, it has heavier tails and can be multimodal depending on how concentrated the data are near rounding thresholds."
   "keywords": [
        "Bayesian inference",
        "rounded data",
        "posterior distribution",
        "truncated likelihood",
        "noninformative prior"
  ],
  "source": "Bayesian Data Analysis, 3rd Edition (Gelman et al., 2013), Exercise 5.9, Chapter 5",
  }
},
  {
    "index": "F3.678",
    "topic": "Logistic Regression",
    "difficulty": "advanced",
    "problem": "A logistic regression model is fitted using a binary predictor: whether an applicant had any type of honors listed on their resume (e.g., 'employee of the month'). The model is loge(p / (1 - p)) = -2.4998 + 0.8668 × honors. (a) What is the callback probability if a resume has no honors? (b) What is the probability if the resume includes honors?",
    "solution_steps": [
      "(a) When honors = 0, the log-odds becomes -2.4998.",
      "Convert log-odds to probability: p = e^(-2.4998) / (1 + e^(-2.4998)) ≈ 0.076.",
      "(b) When honors = 1, the log-odds becomes -2.4998 + 0.8668 = -1.6330.",
      "Then, p = e^(-1.6330) / (1 + e^(-1.6330)) ≈ 0.163."
    ],
    "conclusion": "Including honors on a resume increases the probability of receiving a callback from approximately 7.6% to 16.3%, more than doubling the odds.",
    "explanation": "This problem demonstrates how logistic regression is used to model binary outcomes. The presence of honors is associated with an increase in callback probability, highlighting how predictors affect outcomes on a log-odds scale.",
    "keywords": [
      "logistic regression",
      "log-odds",
      "binary predictor",
      "resume study"
    ],
    "source": "OpenIntro Statistics, 4th Edition (2019), Example 9.31, Chapter 9, Pages 374"
  },
  {
    "index": "F3.789",
    "topic": "Bayesian Statistics",
    "difficulty": "advanced",
    "problem": "Assume a Gamma(4, 5) prior and observe 15 successes and 10 failures. Compute the posterior distribution parameters and posterior mean.",
    "solution_steps": [
      "Step 1: Prior distribution is Gamma(4, 5).",
      "Step 2: Observed data: 15 successes and 10 failures.",
      "Step 3: Update the prior with observed data: Posterior distribution = Gamma(4 + 15, 5 + 10) = Gamma(19, 15).",
      "Step 4: Compute the posterior mean: Mean = α / β = 19 / 15 ≈ 1.267.",
      "Step 5: Compute the posterior variance: Variance = α / β^2 = 19 / 15^2 = 19 / 225 ≈ 0.084.",
      "Step 6: Thus, the posterior distribution is Gamma(19, 15) and the posterior mean is approximately 1.267."
    ],
    "conclusion": "The posterior distribution is Gamma(19, 15) and the posterior mean is approximately 1.267.",
    "explanation": "The Gamma distribution is updated with observed data to yield the posterior distribution parameters.",
    "keywords": [
      "Bayesian statistics",
      "Gamma distribution",
      "posterior distribution",
      "posterior mean",
      "posterior variance"
    ],
    "source": "Understanding Advanced Statistical Methods by Westfall & Henning"
  },
  {
    "index": "F3.890",
    "topic": "Bayesian Inference – Posterior Analysis",
    "difficulty": "advanced",
    "problem": "Suppose you are analyzing a clinical trial for a new drug. Historical data suggest that the probability of a patient responding to the standard treatment (prior probability) is about 0.2. In a new trial with the experimental drug, you observe that out of 30 patients, 10 respond positively. Assume a Beta(2,8) prior distribution for the response probability θ (where Beta(2,8) reflects the historical data). The likelihood is binomial: 10 successes out of 30 trials. Tasks: 1. Calculate the posterior distribution for θ. 2. Find the posterior mean and posterior mode. 3. Construct a 95% credible interval for θ.",
    "solution_steps": [
      "Step 1: Identify the prior distribution. Prior = Beta(2, 8). This reflects prior belief about response probability θ.",
      "Step 2: Extract data from the trial. Number of successes = 10, Number of failures = 20 (since 30 total patients - 10 successes), Likelihood = Binomial(30, θ)",
      "Step 3: Compute the posterior distribution. Posterior = Beta(2 + 10, 8 + 20) = Beta(12, 28)",
      "Step 4: Compute the posterior mean. Mean = α / (α + β) = 12 / (12 + 28) = 0.3",
      "Step 5: Compute the posterior mode. Mode = (α - 1) / (α + β - 2) = (12 - 1) / (12 + 28 - 2) = 11 / 38 ≈ 0.289",
      "Step 6: Compute a 95% credible interval. Use Beta(12, 28) quantiles: Lower bound ≈ 0.17, Upper bound ≈ 0.43. Alternatively, use normal approximation: Variance ≈ (12 * 28) / [(40)^2 * 41] ≈ 0.0042, Std dev ≈ √0.0042 ≈ 0.065, Interval = 0.3 ± 1.96 × 0.065 ≈ (0.173, 0.427)"
    ],
    "conclusion": "The posterior distribution for the response probability θ is Beta(12, 28). The posterior mean is 0.3 and the mode is approximately 0.289. A 95% credible interval for θ is approximately (0.17, 0.43).",
    "keywords": [
      "Bayesian inference",
      "posterior distribution",
      "Beta distribution",
      "credible interval",
      "posterior mean",
      "posterior mode"
    ],
    "source": "Understanding Advanced Statistical Methods by Westfall & Henning (Routledge), Chapter on Bayesian Inference"
  },




  {
    "index": "F3.901",
    "topic": "Logistic Regression – Binary Classification",
    "difficulty": "advanced",
    "problem": "Suppose you are analyzing data from a medical study where the outcome is whether a patient develops a certain disease (yes/no), and the predictors are age (in years) and cholesterol level (mg/dL). The following logistic regression output is obtained: Variable: Intercept, Coefficient: -6.5, Standard Error: 1.2, p-value: <0.001; Variable: Age, Coefficient: 0.04, Standard Error: 0.01, p-value: <0.001; Variable: Cholesterol, Coefficient: 0.02, Standard Error: 0.005, p-value: <0.001. Tasks: 1. Write the equation of the logistic regression model. 2. For a patient aged 50 with a cholesterol level of 200 mg/dL, calculate the log-odds of developing the disease. 3. Calculate the predicted probability of developing the disease for this patient. 4. Interpret the coefficients for age and cholesterol.",
    "solution_steps": [
      "Step 1: Write the logistic regression model. log(p / (1 - p)) = -6.5 + 0.04 * Age + 0.02 * Cholesterol",
      "Step 2: Plug in Age = 50 and Cholesterol = 200 into the model. log(p / (1 - p)) = -6.5 + 0.04 * 50 + 0.02 * 200 = -6.5 + 2 + 4 = -0.5",
      "Step 3: Convert log-odds to probability. p = exp(-0.5) / (1 + exp(-0.5)) ≈ 0.6065 / (1 + 0.6065) ≈ 0.6065 / 1.6065 ≈ 0.377",
      "Step 4: Interpret the coefficients. Age: Each additional year of age increases the log-odds of developing the disease by 0.04, holding cholesterol constant. Cholesterol: Each additional mg/dL of cholesterol increases the log-odds of disease by 0.02, holding age constant."
    ],
    "conclusion": "The logistic regression model estimates a log-odds of -0.5 for a 50-year-old with cholesterol of 200 mg/dL, corresponding to a 37.7% probability of developing the disease. The positive coefficients for age and cholesterol indicate higher risk with increasing values of each.",
    "keywords": [
      "logistic regression",
        "binary classification",
        "log-odds",
        "probability prediction",
        "interpretation of coefficients"
      ],
      "source": "Understanding Advanced Statistical Methods by Westfall & Henning (Routledge), Chapter on Logistic Regression"
    }
  ],
```