Practice Worksheet 1: Foundations of Probability

Core Concepts Review

1. Probability Density/Mass Functions (PDF/PMF)

- **Discrete (PMF):** $p(x) = P(X = x)$ [1]. Must satisfy $p(x) \geq 0$ and $\sum p(x) = 1$.
- **Continuous (PDF):** $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x)dx = 1$ [1].
- **Finding Probabilities:** $P(a \leq X \leq b) = \int_a^b f(x)dx$ [1].

2. Cumulative Distribution Function (CDF) and Percentiles

- **Definition:** $F(t) = P(X \leq t)$ [2].
- **Percentiles:** The $100 * u$-th percentile is the value y_u that solves $P(Y \leq y_u) = u$ [3]. To calculate it, find the inverse CDF: $y_u = F^{-1}(u)$ [3].

3. Expected Value ($E(X)$) and Variance

- **Expected Value (Mean):** $E(X) = \int_{-\infty}^{\infty} xf(x)dx$ (Continuous) or $\sum xp(x)$ (Discrete) [4].
- **Function of a RV:** $E(h(X)) = \int_{-\infty}^{\infty} h(x)f(x)dx$ [5].
- **Variance Formula:** $var(X) = E(X^2) - [E(X)]^2$ [5].
- **Linearity Properties:** $E(aX_1 + bX_2 + c) = aE(X_1) + bE(X_2) + c$ [6].

Practice Problems

Problem 1: Let X be a continuous random variable with density $f(x) = \frac{x}{2}\exp(-x^2/4)1(x \geq 0)$ [7]. Derive $F(x)$, the cumulative distribution function of X [7].

Problem 2: Suppose the random variable X has CDF given by $F_X(t) = 1 - (1 - t^2)^{\theta-1}$ for $0 \leq t \leq 1$ [8]. If $\theta = 2$, calculate the 75th percentile of X [8].

Problem 3: Let $X \sim Unif(-1, 3)$ and define a new random variable $Y = X^2 + 3$ [9]. Derive the cumulative distribution function of Y [9].

Problem 4: Using the random variables from Problem 3, compute $var(Y)$ [9].

Problem 5: The random variable Y has an Exponential distribution with parameter $\mu > 0$, meaning its PDF is $f(y) = \frac{1}{\mu}\exp(-y/\mu)1(y \geq 0)$ [10]. Prove that $var(Y) = \mu^2$ using the variance definition [10, 11]. *Hint:* $\int_0^{\infty} x^n \exp(-x/\mu)dx = n! \cdot \mu^{n+1}$ *[10].*

Problem 6: Let X be the number of successful freethrows made in two attempts, with PMF: $p(0) = 0.25$, $p(1) = 0.5$, and $p(2) = 0.25$ [2]. Calculate $E(X^2)$ and $var(X)$ [12].

Problem 7: Suppose $X_1, \ldots, X_n$ are iid $Bernoulli(\theta)$ [13]. What is the exact distribution of $\sum_{i=1}^n X_i$? Provide the name and its parameters [14].

Problem 8: Let $Y \sim Exp(\mu)$ [10]. Is the statement $P(0 \leq Y \leq 2) = P(0 \leq Y < 2)$ TRUE or FALSE? Explain why [15].

Problem 9: For any two random variables X_1 and X_2 and fixed constants a, b, expand $var(aX_1 + bX_2)$ [6]. Under what specific condition does this simplify to $a^2var(X_1) + b^2var(X_2)$ [6]?

Problem 10: Using the inversion method, write an expression for a random variable that has the same distribution as X from Problem 2 (where $F_X(t) = 1 - (1 - t^2)^{\theta-1}$) [8].

Practice Worksheet 2: Named Distributions and R Coding

Core Concepts Review

1. Uniform Distribution: $Unif(\alpha_1, \alpha_2)$

- **PDF:** $f(x) = \frac{1}{\alpha_2 - \alpha_1}$ for $x \in [\alpha_1, \alpha_2]$.
- **Mean & Variance:** $E(X) = \frac{\alpha_1 + \alpha_2}{2}$, $var(X) = \frac{(\alpha_2 - \alpha_1)^2}{12}$.
- **Transformation:** If $U \sim Unif(0, 1)$, then $a + (b - a)U \sim Unif(a, b)$.

2. Bernoulli and Binomial Distributions

- **Bernoulli(θ):** Models a single success/failure. $E(X) = \theta$, $var(X) = \theta(1 - \theta)$.
- **Binomial(n, θ):** Sum of n iid Bernoulli(θ) trials. $E(X) = n\theta$, $var(X) = n\theta(1 - \theta)$.
- **Simulation Trick:** $1(U \leq \theta) \sim Bern(\theta)$ where $U \sim Unif(0, 1)$.

3. Normal Distribution: $N(\mu, \sigma^2)$

- **Mean & Variance:** $E(X) = \mu$, $var(X) = \sigma^2$ (Note: R functions use σ, not σ^2).
- **Transformation:** If $Z \sim N(0, 1)$, then $\mu + \sigma Z \sim N(\mu, \sigma^2)$.
- **Linear Combinations:** If $X_i \sim N(\mu_i, \sigma_i^2)$ are independent, $\sum b_i X_i \sim N(\sum b_i \mu_i, \sum b_i^2 \sigma_i^2)$.

4. Core R Functions

- `r[name](n, ...)`: Generates n random realizations.
- `p[name](q, ...)`: Computes the CDF, $P(X \leq q)$.
- `q[name](p, ...)`: Computes the inverse CDF (percentile) for probability p.
- `d[name](x, ...)`: Computes the density or probability mass at x.

Practice Problems

Problem 1: Let $X \sim Unif(-2, 4)$. Calculate the exact probability $P(X < 0)$.

Problem 2: Suppose you only have access to the function `runif(n=1)`, which generates a single $Unif(0, 1)$ variable. Write a single line of R code that transforms this standard uniform into a realization of $Y \sim Unif(5, 10)$.

Problem 3: Let $X \sim Binom(15, 0.4)$. Express X mathematically as a sum of simpler random variables, and state the exact expected value and variance of X.

Problem 4: Fill in the missing R code labeled with `?????` so that $Y \sim Binomial(n = 10, \theta = 0.2)$.

```
X <- 1*(rnorm(10) < ?????)
Y <- sum(X)
```

Problem 5: The random variable X has a Normal distribution with mean 10 and variance 16. Write the exact R code using the `pnorm` function to compute $P(X > 12)$.

Problem 6: Let $Z \sim N(0, 1)$. Define a new random variable $Y = 3 - 2Z$. What is the exact distribution of Y? Provide the distribution name, mean, and variance.

Problem 7: Without using the `rbinom` function, write a single line of R code that generates a realization of 50 independent copies of a $Bernoulli(0.75)$ random variable.

Problem 8: Suppose X_1, X_2, X_3 are independent and identically distributed $N(5, 4)$ random variables. Find the exact distribution of $Y = X_1 + X_2 + X_3$, including its parameters.

Problem 9: You are given that `pnorm(2.5)` ≈ 0.9938. Using the symmetry of the standard normal distribution, what is the value of `pnorm(-2.5)`?

Problem 10: Write the R code to efficiently generate a matrix called `Y_mat` with 1000 rows and 15 columns, where every entry is an independent realization of a $N(\mu = 8, \sigma^2 = 9)$ random variable.

Practice Worksheet 3: Estimators vs. Estimates

Core Concepts Review

The Big Picture: Why do we care?

In statistics, there is always a "hidden truth" about a population (like the true average height of everyone on Earth). We call this truth a **parameter** (often denoted by Greek letters like μ or σ^2). Because we cannot measure everyone on Earth, we take a smaller sample.

We use formulas on our sample data to *guess* the hidden truth. The concepts of "Estimators" and "Estimates" are just the formal ways we talk about that guessing process before and after we actually collect the data.

1. Estimators (The Formula / The Strategy)

An **estimator** is the mathematical formula or rule you plan to use *before* you collect any data.

- Because you haven't collected the data yet, the result is still a mystery. Therefore, an estimator is a **random variable**.
- We use **capital letters** to represent estimators (and the random variables they are built from).
- **Sample Mean Estimator:** $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- **Sample Variance Estimator:** $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

2. Estimates (The Actual Number / The Result)

An **estimate** is the actual, hard number you calculate *after* you have collected your data and plugged it into your estimator formula.

- Because the data is already collected, there is no more mystery. An estimate is **not random**; it is just a fixed number.
- We use **lowercase letters** to represent estimates (and the observed data).
- **Sample Mean Estimate:** $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- **Sample Variance Estimate:** $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

3. Properties of Good Estimators

We want our estimators to be accurate on average. If the expected value of an estimator equals the true hidden parameter, we call it **unbiased**.

- $E(\bar{X}) = \mu$ (The sample mean is an unbiased estimator of the true mean).
- $E(S^2) = \sigma^2$ (The sample variance is an unbiased estimator of the true variance. This is exactly why we divide by $n-1$ instead of n!).
- $var(\bar{X}) = \frac{\sigma^2}{n}$ (As your sample size n gets larger, the variance of your sample mean shrinks, meaning your guesses get tighter and more accurate).

Practice Problems

Problem 1: Explain the fundamental difference between an estimator and an estimate in one sentence.

Problem 2: You are planning to measure the exact weights of 15 randomly selected apples. Let $X_1, \ldots, X_{15}$ represent these yet-to-be-observed weights. Is $\frac{1}{15}\sum_{i=1}^{15} X_i$ an estimator or an estimate? Is it a random variable?

Problem 3: You weigh the 15 apples and calculate an average weight of 142 grams. Is 142 an estimator or an estimate? Is it a random variable?

Problem 4: Let $X_1, \ldots, X_n$ be independent copies of a random variable X with a true population mean of $E(X) = 10$. What is the expected value of the sample mean, $E(\bar{X})$?

Problem 5: Let $X_1, \ldots, X_n$ be independent copies of a random variable X with a true population variance of $var(X) = 25$. If your sample size is $n = 100$, what is the variance of the sample mean, $var(\bar{X})$?

Problem 6: Suppose you have a realized sample of data: $x_1 = 2, x_2 = 4, x_3 = 6$. Calculate the sample mean estimate, $\bar{x}$.

Problem 7: Using the exact same realized data from Problem 6 ($x_1 = 2, x_2 = 4, x_3 = 6$), calculate the sample variance estimate, s^2. *Hint: Remember to divide by $n - 1$.*

Problem 8: The random variable S^2 is the sample variance estimator. If the data comes from a population with a true variance of $\sigma^2 = 14$, what is $E(S^2)$?

Problem 9: In R, suppose you have a vector of 50 realized data points stored in a variable called `my_data`. Write the two specific R functions you would use to calculate the estimates $\bar{x}$ and s^2.

Problem 10: As your sample size n approaches infinity, what happens to the value of $var(\bar{X})$? Conceptually, what does this tell you about the accuracy of your sample mean as you collect more data?

Practice Worksheet 4: The Inversion Method

Core Concepts Review

The Goal

Computers naturally know how to generate random numbers uniformly between 0 and 1 using a standard uniform generator (like `runif()` in R). The Inversion Method is a mathematical trick that lets us bend those uniform numbers into *any* shape (distribution) we want, as long as we know the target's Cumulative Distribution Function (CDF).

The Theorem

If $U \sim Unif(0,1)$ and $F(x)$ is the cumulative distribution function (CDF) of our target distribution, then the random variable $X = F^{-1}(U)$ has the exact same distribution as $F(x)$.

The 4-Step Process

1. **Find the CDF:** If you are only given the PDF $f(x)$, integrate it to find $F(x) = \int_{-\infty}^{x} f(t)dt$.
2. **Set $u = F(x)$:** Create an algebraic equation.
3. **Solve for x:** Isolate x on one side to find the inverse function, $x = F^{-1}(u)$.
4. **Plug in U:** Replace the lowercase u with the random variable U. Your final generator is $X = F^{-1}(U)$.

In R

To simulate this, simply draw standard uniform variables using `U <- runif(n)`, and then plug that `U` directly into the F^{-1} formula you derived.

Practice Problems

Problem 1: Explain the primary purpose of the Inversion Method in one sentence.

Problem 2: Let $X \sim Exp(\mu)$ with CDF $F(x) = 1 - \exp(-x/\mu)$ for $x \geq 0$. Apply the inversion method to derive an expression for X in terms of $U \sim Unif(0,1)$.

Problem 3: Write an R function `rexp_custom(n, mu)` that uses your mathematical result from Problem 2 to generate `n` independent copies of $X \sim Exp(\mu)$. Do not use R's built-in `rexp` function.

Problem 4: The continuous random variable X has CDF $F(x) = x^2$ for $x \in [1]$. Find the inversion method formula to generate a realization of X from $U \sim Unif(0,1)$.

Problem 5: The triangular distribution has density $f(x) = 2(1-x)$ for $x \in [1]$. Integrating this yields the CDF $F(x) = 2x - x^2$. Derive $F^{-1}(u)$ for $u \in (0,1)$. *Hint: You will need to use the quadratic formula or complete the square.*

Problem 6: In R, write a single line of code that generates 50 realizations of the triangular random variable from Problem 5, assuming `U <- runif(50)` has already been run.

Problem 7: Suppose a random variable Y has density $f(y) = (\lambda+1)y^{\lambda}$ for $0 \leq y \leq 1$ and $\lambda > 0$. Derive the CDF, $F(y)$.

Problem 8: Using your CDF from Problem 7, derive the inversion method generator for Y.

Problem 9: Suppose X has the CDF $F(t) = 1 - (1-t^2)^{\theta-1}$ for $0 \leq t \leq 1$, where $\theta > 1$. Using the inversion method, write an expression for X depending only on $U \sim Unif(0,1)$.

Problem 10: When checking if your inversion method worked in R, you generate 10,000 realizations and plot them against the theoretical quantiles in a QQ-plot. If you already have your sequence of probabilities stored in `probs <- ppoints(10000)`, how do you calculate the theoretical quantiles (the x-axis values) without using any built-in 'q' functions like `qnorm` or `qexp`?

Practice Worksheet 5: Monte Carlo Simulations and Box-Muller

Core Concepts Review

1. The Box-Muller Method

The Box-Muller method is a clever mathematical trick to turn two independent Uniform random variables into two independent Standard Normal random variables.

- **The Theorem:** If $U_1, U_2 \sim Unif(0,1)$ are independent, then:

$$X_1 = \sqrt{-2\log(U_1)}\cos(2\pi U_2)$$
$$X_2 = \sqrt{-2\log(U_1)}\sin(2\pi U_2)$$

  Both X_1 and X_2 are exactly independent $N(0,1)$ random variables.

- **Transforming to any Normal:** Once you have $Z \sim N(0,1)$ from Box-Muller, you can shift and scale it to any Normal distribution $Y \sim N(\mu, \sigma^2)$ using $Y = \mu + \sigma Z$.

- **The CLT Alternative:** If you cannot use sine or cosine, you can approximate a $N(0,1)$ variable by summing 12 independent $Unif(0,1)$ variables and subtracting 6: $Z \approx \sum_{i=1}^{12} U_i - 6$.

2. Monte Carlo Simulations

Monte Carlo methods use repeated random sampling to estimate numerical results. Because we control the "truth" in a simulation, we can use it to test how well statistical procedures work.

- **Law of Large Numbers (LLN):** If you generate thousands of independent realizations of a random variable, their sample average will converge to the true expected value.

- **Estimating Probabilities:** Probabilities are just expected values of indicator functions. To estimate $P(X > 5)$, you simulate X thousands of times and calculate the proportion of times $X > 5$ occurred (e.g., `mean(X > 5)` in R).

- **Reproducibility:** Always use `set.seed(number)` before a simulation so your random draws are exactly repeatable.

Practice Problems

Problem 1: Suppose $U_1 = 0.5$ and $U_2 = 0.25$ are realizations of two independent standard uniform random variables. Use the Box-Muller formulas to calculate the resulting realizations of X_1 and X_2.

Problem 2: You have used the Box-Muller method to generate $Z \sim N(0,1)$. Write the mathematical formula to transform Z into a realization of $W \sim N(15, 9)$.

Problem 3: Why does the alternative approximation formula $\sum_{i=1}^{12} U_i - 6$ mathematically yield a mean of 0 and variance of 1? *Hint: Think about the mean and variance of a single $Unif(0,1)$ variable.*

Problem 4: Write R code using the Box-Muller method to generate a vector called `Z_vals` containing 10,000 independent $N(0,1)$ realizations. Do not use the `rnorm()` function.

Problem 5: Explain in one sentence the primary purpose of using `set.seed()` before running a Monte Carlo simulation.

Problem 6: Suppose you want to estimate the true expected value of $|X - 4|$ where $X \sim N(3, 16)$. Write an R script that uses a Monte Carlo simulation with 5,000 replications to estimate this value.

Problem 7: In R, you have generated a vector `t_stats` containing 10,000 simulated test statistics. You want to reject the null hypothesis if the absolute value of the test statistic is strictly greater than 2.1. Write a single line of R code to estimate the probability of rejection.

Problem 8: A student runs a Monte Carlo simulation to estimate a Type I error probability and gets an estimate of 0.0532. The true significance level was $\alpha = 0.05$. Does this mean the test is broken? Explain what causes this discrepancy.

Problem 9: Suppose $X_1, \ldots, X_n$ are iid $Exp(\mu)$. You are asked to perform a simulation study to see if the sample mean $\bar{X}$ is unbiased. Outline the 3 main coding steps you would take to prove $E(\bar{X}) \approx \mu$ via simulation.

Problem 10: When creating a Monte Carlo simulation in R, why is it generally preferred to use matrix operations and the `apply()` function instead of writing a `for` loop?

Practice Worksheet 6: Central Limit Theorem and QQ-Plots

Core Concepts Review

1. The Central Limit Theorem (CLT)

The Central Limit Theorem is one of the most powerful rules in statistics. It states that if you take independent and identically distributed (iid) samples from *any* distribution (as long as it has a finite mean μ and variance σ^2), the distribution of the **sample mean ($\bar{X}$)** will eventually look like a Normal curve as your sample size n gets larger.

- **The Approximation:** For large n, $\bar{X} \approx N(\mu, \frac{\sigma^2}{n})$.
- **Why it matters:** This means we can use Normal distribution math (like Z-scores and `pnorm`) to find probabilities about sample means, even if the original data was Exponential, Uniform, or Binomial!
- **The Catch:** If the original distribution is highly skewed or weird, you need a much larger n for the Normal approximation to become accurate.

2. Reading QQ-Plots (Quantile-Quantile Plots)

A QQ-plot is a visual test to see if your data follows a specific theoretical distribution (usually the Normal distribution). We plot the theoretical percentiles on the x-axis and your actual data percentiles on the y-axis.

- **The Perfect Match:** If your data perfectly follows the theoretical distribution, every data percentile will exactly equal the theoretical percentile. The points will form a perfect diagonal line: $y = x$.
- **Reading the Tails:**
  - If points on the far right fall *below* the line, your data's right tail is *lighter* (shorter) than the theoretical distribution.
  - If points on the far left fall *above* the line, your data's left tail is *lighter* than the theoretical distribution.
  - (If they fall outside the line in the opposite directions, the tails are *heavier*).

3. QQ-Plots in R

To build a QQ-plot manually in R, you generally follow these steps:

1. Create an evenly spaced sequence of probabilities using `probs <- ppoints(length(data))`.
2. Calculate the theoretical quantiles using the inverse CDF (e.g., `qnorm(probs, mean, sd)`).
3. Calculate the data quantiles using `quantile(data, probs)`.
4. Plot them against each other using `plot(theoretical, data)`.
5. Add the $y = x$ reference line using `abline(0, 1)` (or `abline(a=0, b=1)`).

Practice Problems

Problem 1: Explain the Central Limit Theorem in your own words, specifically mentioning what happens to the sample mean as n approaches infinity.

Problem 2: Suppose $X_1, \ldots, X_{50}$ are iid $Unif(1, 10)$. Using the CLT, what is the approximate Normal distribution of the sample mean, $\bar{X}$? Provide the mean and variance of this approximate distribution.

Problem 3: You simulate the sample mean $\bar{X}$ of an Exponential distribution with $n = 2$ and create a Normal QQ-plot. The plot is heavily curved and does not match the $y = x$ line. Does this mean the Central Limit Theorem is false? Explain.

Problem 4: In R, what does the `ppoints(n)` function do, and why is it specifically useful when constructing QQ-plots?

Problem 5: You are creating a Normal QQ-plot and have already run `probs <- ppoints(500)`. Write the exact R code to calculate the theoretical quantiles for a Normal distribution with a mean of 15 and a standard deviation of 4.

Problem 6: Write the R code to compute the sample quantiles of a vector of data called `my_sim_data`, using the `probs` vector from Problem 5.

Problem 7: After plotting your theoretical quantiles on the x-axis and data quantiles on the y-axis, write the exact R code required to add the 1-to-1 reference line to the plot.

Problem 8: Looking at a Normal QQ-plot, you notice that the points on the far right side of the graph dip noticeably *below* the $y = x$ reference line. What does this tell you about the right tail of your data compared to a true Normal distribution?

Problem 9: Suppose you are testing if a test statistic T follows a t-distribution with $n - 1$ degrees of freedom, rather than a Normal distribution. If your probabilities are stored in `probs`, what R function would you use to calculate the theoretical quantiles for the x-axis?

Problem 10: A classmate claims that the CLT proves that if you collect enough data points, the *population data itself* will eventually turn into a Normal distribution. Explain why your classmate is dangerously incorrect.

Practice Worksheet 7: Confidence Intervals and Coverage Probability

Core Concepts Review

The Big Picture: What is a Confidence Interval?

When we take a sample, we calculate an estimate (like the sample mean, $\bar{x}$) to guess the true population parameter (like the true mean, μ). However, we know our single guess won't be perfectly accurate.

A **confidence interval** gives us a margin of error around our guess. Instead of saying "the mean is exactly 65," we say "we are highly confident the true mean is between 63 and 67."

1. The Most Common Mistake: Interpreting the Interval

It is crucial to understand what is random and what is fixed:

- **The True Parameter (μ) is FIXED.** It is a single, hidden truth. It is not moving around. It is not a random variable.
- **The Interval is RANDOM.** Because the interval is calculated from your random sample data, the *interval itself* changes every time you take a new sample.

The Trap: Once you calculate a specific interval with actual data (e.g., $[111.8, 116.7]$), it is wrong to say "There is a 95% probability that the true mean is in this interval." Why? Because the true mean is a fixed number, and your calculated interval is fixed numbers. The mean is either in there (100% probability) or it isn't (0% probability).

The Correct Interpretation: "We are 95% confident that the true mean is between 111.8 and 116.7." This means that if we repeated our experiment infinitely many times and built a new interval each time, 95% of those generated intervals would successfully cover the true, fixed mean.

2. Coverage Probability

Coverage probability is the "before data is collected" perspective. Before you run your experiment, the coverage probability is the exact likelihood (usually $1 - \alpha$, like 0.95) that your *yet-to-be-calculated* random interval will capture the true parameter.

3. The One-Sample t-Interval for the Mean

When estimating a mean μ from a Normal distribution (or any distribution if n is large), we use the sample mean $\bar{X}$ and sample standard deviation S. Because we are guessing S instead of knowing the true σ, the math requires us to use the heavier-tailed **t-distribution** instead of the Normal distribution.

The random $100(1 - \alpha)\%$ confidence interval for μ is:

$$\bar{X} \pm t_{1-\alpha/2,n-1} \frac{S}{\sqrt{n}}$$

Where $t_{1-\alpha/2,n-1}$ is the percentile from the t-distribution with $n - 1$ degrees of freedom.

Practice Problems

Problem 1: You calculate a 95% confidence interval for the average points scored by the Timberwolves and get $[111.8, 116.7]$. Your friend says, "There is a 95% chance that the true average is exactly 115." Explain why this statement is mathematically incorrect.

Problem 2: Using the same interval from Problem 1, another friend says, "There is a 95% probability that the true average is between 111.8 and 116.7." Is this interpretation correct or incorrect? Explain why based on what is "fixed" versus what is "random."

Problem 3: Write out the exact, correct layman's interpretation of the confidence interval $[111.8, 116.7]$ from Problem 1.

Problem 4: In the confidence interval formula $\bar{X} \pm t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}}$, what does the piece $\frac{S}{\sqrt{n}}$ represent?

Problem 5: You want to construct a 99% confidence interval ($\alpha = 0.01$) from a sample of size $n = 20$. Write the exact R code you would use to find the appropriate t-distribution multiplier, $t_{1-\alpha/2,n-1}$.

Problem 6: Suppose your sample mean is $\bar{x} = 64.5$, your sample standard deviation is $s = 5.0$, your sample size is $n = 100$, and your t-multiplier is 1.98. Calculate the lower and upper bounds of this confidence interval.

Problem 7: If you decide you want to be 99% confident instead of 95% confident, but you keep the exact same data, will your new confidence interval be wider or narrower? Why?

Problem 8: If you collect a much larger sample of data (increasing n from 30 to 300), what will mathematically happen to the width of your confidence interval?

Problem 9: When building a confidence interval for the mean, why do we generally use the t-distribution ($qt()$ in R) instead of the standard Normal distribution ($qnorm()$ in R)?

Problem 10: You want to run a simulation study to prove that the t-interval formula actually has a 95% coverage probability. You generate a matrix of 10,000 confidence interval lower bounds (`LBs`) and upper bounds (`UBs`), and you know the true mean is $\mu = 68$. Write a single line of R code that calculates the proportion of these 10,000 intervals that successfully captured the true mean.

Practice Worksheet 8: One-Sample CIs for the Mean (t-distribution)

Core Concepts Review

The Big Picture: Why do we need a new distribution?

When we want to build a confidence interval for a population mean (μ), we start with our sample mean ($\bar{X}$). From the Central Limit Theorem, we know that if we standardize $\bar{X}$ using the true population standard deviation (σ), it perfectly follows a Standard Normal distribution $N(0,1)$:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

The Problem: In reality, we almost never know the true population standard deviation (σ). **The Solution:** We have to guess it using our sample standard deviation (S).

But when we swap the fixed, constant σ for a random, fluctuating S, we introduce extra uncertainty into our formula. Because of this extra uncertainty, the result no longer perfectly fits a Normal distribution. Instead, it follows a **t-distribution**:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim T_{n-1}$$

The t-distribution (T_{n-1})

- **Degrees of Freedom:** The t-distribution relies on a parameter called "degrees of freedom", which is $n-1$ (your sample size minus one).
- **Shape:** It looks like a Normal curve, but it has "heavier" or "thicker" tails. This mathematically accounts for the extra uncertainty of guessing σ.
- **Convergence:** As your sample size n gets very large (e.g., $n > 500$), your guess S becomes so accurate that the t-distribution looks virtually identical to a Normal distribution.

The Confidence Interval Formula

By rearranging the t-distribution formula, we get the random $100(1-\alpha)\%$ confidence interval for μ:

$$\left[\bar{X} - t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}}, \quad \bar{X} + t_{1-\alpha/2,n-1}\frac{S}{\sqrt{n}}\right]$$

- **Exact vs. Approximate:** If your original data comes from a Normal distribution, this formula has *exactly* a $1-\alpha$ coverage probability. If your data comes from any other distribution, this formula is only *approximate*, but becomes highly accurate as n increases thanks to the Central Limit Theorem.
- **In R:** You can compute this entire interval instantly using `t.test(data, conf.level = 0.95)$conf.int`.

Practice Problems

Problem 1: Explain in your own words why we must use the t-distribution instead of the Normal distribution when building a confidence interval for the mean from a set of data.

Problem 2: Suppose you are calculating a 90% confidence interval ($\alpha = 0.10$) from a sample of 25 observations. Write the exact R code to find the t-multiplier, $t_{1-\alpha/2,n-1}$.

Problem 3: You have a dataset of 40 observations. The sample mean is 120 and the sample standard deviation is 15. Assuming the t-multiplier is approximately 2.02, calculate the upper and lower bounds of this confidence interval.

Problem 4: You calculate a confidence interval for the mean and get $[10.5, 14.2]$. A classmate claims, "With 95% confidence, the value 0 is a plausible true mean based on our sample." Is this TRUE or FALSE? Explain.

Problem 5: Let $X_1, \ldots, X_n$ be an iid sample. If this data comes from an Exponential distribution, is the t-interval formula exact or approximate? What mathematical theorem makes it acceptable to use anyway if n is large?

Problem 6: Suppose your data is $X_1, \ldots, X_{20}$ iid $N(\mu, \sigma^2)$. Is the coverage probability of the t-interval exactly $1 - \alpha$, or approximately $1 - \alpha$?

Problem 7: In R, you have a vector of data called `temp_data`. Write the single line of code that uses a built-in R function to calculate the 99% confidence interval for the mean.

Problem 8: Look at the confidence interval formula: $\bar{X} \pm t_{1-\alpha/2, n-1} \frac{S}{\sqrt{n}}$. If you increase your sample size n from 20 to 200, but $\bar{X}$ and S remain exactly the same, what will happen to the overall width of the interval?

Problem 9: Again looking at the formula, if you decide you want to be 99% confident instead of 90% confident, what specific piece of the formula changes, and does the interval get wider or narrower?

Problem 10: When running a simulation study to prove this interval works, you generate 10,000 realizations of the test statistic $T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$. To visually check if T truly follows a T_{n-1} distribution, you decide to make a QQ-plot. What specific R function must you use to generate the theoretical quantiles for the x-axis?

Practice Worksheet 9: CIs for Bernoulli Success Probability

Core Concepts Review

The Big Picture: Why are there four different intervals?

When we have n binary (0 or 1) observations, we model them as independent $Bernoulli(\theta)$ random variables. We want to estimate the true success probability, θ.

Using the Central Limit Theorem (CLT), we know the sample proportion $\bar{X}$ is approximately Normal. Standardizing it gives us:

$$\frac{\bar{X} - \theta}{\sqrt{\theta(1-\theta)/n}} \approx N(0,1)$$

To build a $100(1-\alpha)\%$ confidence interval, we need to solve this double inequality for θ:

$$-z_{1-\alpha/2} < \frac{\bar{X} - \theta}{\sqrt{\theta(1-\theta)/n}} < z_{1-\alpha/2}$$

The Problem: Notice that the unknown parameter θ is in both the numerator *and* the denominator inside the square root. The four different intervals we use are simply four different mathematical strategies to handle this annoying algebra problem.

1. The Score Interval (The "Algebraic Beast")

Instead of taking a shortcut, you can square both sides of the inequality and use the quadratic formula to perfectly isolate θ. This results in a massive, ugly formula.

- **Pros:** It is highly accurate and performs the best in simulation studies (narrow but maintains coverage).
- **In R:** `prop.test(x = n*xbar, n = n, conf.level = 1-alpha, correct = FALSE)$conf.int`

2. The Wald Interval (The "Lazy Shortcut")

Instead of doing the hard algebra, the Wald method just takes a guess. It replaces the unknown θ in the denominator with our best guess, the sample mean $\bar{X}$.

- **Formula:** $\bar{X} \pm z_{1-\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}$
- **Pros/Cons:** Easiest to calculate by hand, but in simulations, it frequently "under-covers" (its actual coverage probability is often lower than the requested $1-\alpha$).

3. The Simple Conservative Interval (The "Worst-Case Scenario")

The mathematical function $\theta(1-\theta)$ hits its absolute maximum possible value when $\theta = 0.5$ (where $0.5 \times 0.5 = 0.25$). The conservative method plugs in this worst-case variance. Since $\sqrt{0.25} = 0.5$, the formula simplifies beautifully.

- **Formula:** $\bar{X} \pm z_{1-\alpha/2}\frac{0.5}{\sqrt{n}}$
- **Pros/Cons:** Guaranteed to never under-cover because it assumes the maximum possible margin of error, but it is often unnecessarily wide.

4. The Clopper-Pearson Interval (The "Exact Method")

This method abandons the Central Limit Theorem and the Normal approximation entirely. It uses exact Binomial probabilities to find the interval, which results in formulas relying on the F-distribution.

- **Pros/Cons:** Guarantees coverage probability of *at least* $1-\alpha$, but can be overly conservative (wide).
- **In R:** `binom.test(x = n*xbar, n = n, conf.level = 1-alpha)$conf.int`

Practice Problems

Problem 1: Explain in your own words why the Wald interval formula uses $\bar{X}(1-\bar{X})$ inside the square root instead of $\theta(1-\theta)$.

Problem 2: You have a sample of $n = 100$ Bernoulli trials and observe a sample mean of $\bar{x} = 0.36$. Using $z_{0.975} \approx 1.96$, calculate the 95% Wald confidence interval by hand.

Problem 3: Using the exact same data from Problem 2 ($n = 100$, $\bar{x} = 0.36$, $z = 1.96$), calculate the 95% Simple Conservative confidence interval by hand.

Problem 4: Compare your answers from Problem 2 and Problem 3. Which interval is wider? Explain mathematically why the Conservative interval will *always* be at least as wide as the Wald interval.

Problem 5: Write the exact R code to compute the 90% Score confidence interval for a dataset of $n = 50$ where you observed 40 successes. *Hint: remember the* `correct=FALSE` *argument.*

Problem 6: Write the exact R code to compute the 99% Clopper-Pearson confidence interval for a dataset of $n = 80$ where the observed sample proportion was $\bar{x} = 0.25$.

Problem 7: In a simulation study comparing these intervals, you see that the Wald interval only successfully captured the true parameter 92% of the time, even though you requested a 95% confidence level. What is the statistical term for this 92%?

Problem 8: Based on the simulation studies shown in lecture, which of the four intervals generally performs the "best" (meaning it reliably hits its target coverage probability while remaining as narrow as possible)?

Problem 9: When dealing with these Bernoulli confidence intervals, why do we use z-percentiles (from the standard Normal distribution) instead of t-percentiles (like we did for estimating a population mean)?

Problem 10: In R, write a boolean expression (that evaluates to TRUE or FALSE) to check if a specific Wald interval, with lower bound `wL` and upper bound `wR`, successfully captured the true parameter `theta`.

Practice Worksheet 10: Type I/II Errors and p-values

Core Concepts Review

The Big Picture: The Decision Matrix

Whenever we run a hypothesis test, we are making a decision based on incomplete information (our sample). Because our sample is random, our decision might be wrong. There are two "truths" (Null H_0 is true, or Alternative H_a is true) and two "decisions" (Reject H_0, or Fail to Reject H_0).

This creates a 2×2 matrix of possible outcomes:

- **Correct Decision 1 ($1 - \alpha$):** H_0 is true, and we correctly fail to reject it.
- **Type I Error (α):** H_0 is true, but we incorrectly reject it. (The "False Alarm"). The probability of this happening is our significance level, α.
- **Type II Error (β):** H_a is true, but we incorrectly fail to reject H_0. (The "Missed Detection").
- **Correct Decision 2 / Power ($1 - \beta$):** H_a is true, and we correctly reject H_0.

Understanding the p-value

The p-value is one of the most misunderstood concepts in statistics.

- **What it is NOT:** The p-value is *not* the probability that the null hypothesis is true. It is also *not* the probability that the alternative hypothesis is true.
- **What it IS (Observed p-value):** Imagine the null hypothesis is an absolute, undeniable fact. The observed p-value is the probability that, if you ran your exact experiment again, you would get a test statistic that is *as extreme or even more extreme* than the one you just got.
- **The Logic:** If the p-value is incredibly small (e.g., 0.001), it means your data is wildly unlikely to happen in a world where H_0 is true. Because your data actually *did* happen, we conclude that the H_0 world must be a lie, so we reject it in favor of H_a.
- **The Rule:** We reject H_0 if the observed p-value is strictly less than our chosen Type I error rate, α.

The "Random" p-value and its Distribution

Before we collect data, the p-value we *will* get is a mystery, making it a random variable.

- **When H_0 is TRUE:** The random p-value perfectly follows a Standard Uniform distribution: $Unif(0, 1)$. This is a beautiful mathematical guarantee! It means that if we set $\alpha = 0.05$, there is exactly a 5% chance our random p-value drops below 0.05, meaning we perfectly control our Type I error rate.
- **When H_a is TRUE:** The p-value no longer follows a uniform distribution. Instead, its probability mass heavily clusters near zero. We *want* it to be near zero so that we can correctly reject H_0.

Practice Problems

Problem 1: You run a test and decide to reject the null hypothesis. Later, an oracle tells you the null hypothesis was actually true all along. What specific type of error did you commit?

Problem 2: You are testing a new drug. The null hypothesis (H_0) is that the drug does nothing. The alternative (H_a) is that the drug cures the disease. Explain what a Type II error represents in the real-world context of this drug.

Problem 3: A classmate runs a t-test and gets a p-value of 0.02. They write in their report: "There is a 2% probability that the null hypothesis is true." Explain why your classmate is entirely incorrect.

Problem 4: Using the exact definition from lecture, write out the correct interpretation of the observed p-value of 0.02 from Problem 3.

Problem 5: You are simulating the Type I error probability of a one-sample t-test. You generate 10,000 p-values under a scenario where H_0 is completely true, and store them in a vector called `pvals`. What exact R distribution should the histogram of `pvals` look like?

Problem 6: Using the `pvals` vector from Problem 5, write a single line of R code that produces a simulation-based estimate of your test's Type I error probability at a significance level of $\alpha = 0.05$.

Problem 7: Suppose you change your simulation so that the alternative hypothesis (H_a) is now strongly true. You generate 10,000 new p-values. Will the histogram of these new p-values still look like a flat $Unif(0, 1)$ distribution? If not, where will the majority of the data be clustered?

Problem 8: The true power of a specific statistical test is 0.85. What is the probability that this test commits a Type II error?

Problem 9: You perform a one-sample t-test with a significance level of $\alpha = 0.01$ and observe a test statistic of $t = 3.5$. The calculated p-value is 0.008. Do you Reject or Fail to Reject H_0?

Problem 10: You run a simulation study for a test statistic that does *not* perfectly follow its assumed distribution. You find that your simulated Type I error probability is 0.065, even though you set $\alpha = 0.05$. What does this tell you about the accuracy of your test's p-values in this specific scenario?

Practice Worksheet 11: Power and Power Curves

Core Concepts Review

The Big Picture: What is Power?

Whenever we run a hypothesis test, we hope to make the correct decision. If the null hypothesis (H_0) is a lie, we want to expose it by rejecting it.

Power is the exact probability that we successfully reject H_0 when the alternative hypothesis (H_a) is actually true. In the real world, a test with high power is highly sensitive—it is very good at detecting a real effect or difference.

Mathematically, Power $= 1 - \beta$ (where β is the probability of a Type II error, or a "Missed Detection").

1. Calculating Power

In our standard one-sample t-test, we reject H_0 if our test statistic T is extreme: $|T| > t_{1-\alpha/2,n-1}$.

Therefore, the formula for power is simply the probability of that event occurring: $P(|T| > t_{1-\alpha/2,n-1})$.

The Catch: When H_0 is true, T perfectly follows a t-distribution. But power only exists in a world where H_a is true (meaning $\mu \neq \mu_0$). In this H_a world, the distribution of T shifts and gets complicated. Because the exact math is messy, we use Monte Carlo simulations to estimate it.

2. Simulating Power in R

To estimate power via simulation, we force H_a to be true, generate data, and see how often our test catches it.

1. **Generate the Data under the TRUTH:** Use the *true* mean (μ) to generate data. (e.g., `x <- rnorm(n, mean = mu, sd = sigma)`).
2. **Test against the NULL:** When calculating the test statistic, you must plug in the *null* mean (μ_0), because the test itself doesn't know the truth! $T = \frac{\bar{X}-\mu_0}{S/\sqrt{n}}$.
3. **Make a Decision:** Check if your p-value is less than α, or if $|T| > t_{1-\alpha/2,n-1}$.
4. **Repeat and Average:** Do this 10,000 times. The proportion of times you correctly rejected H_0 is your estimated power.

3. Power Curves

A power curve is just a line graph that shows how power changes when you tweak the rules of the game.

- **Power vs. True Mean (μ):** As the true mean gets further and further away from the null mean μ_0 (meaning $|\mu - \mu_0|$ increases), the problem becomes "easier" to detect. The power curve will swoop upward toward 1.0 (100% power). If the true mean exactly equals the null mean ($\mu = \mu_0$), the power curve drops down to exactly α (because it becomes a Type I error!).
- **Power vs. Sample Size (n):** As you collect more data (n increases), your estimates get tighter and more accurate. This makes it easier to detect a false H_0. Thus, as n increases, the power curve swoops upward toward 1.0.

Practice Problems

Problem 1: Define statistical "power" in your own words.

Problem 2: If a statistical test has a 15% chance of committing a Type II error, what is the power of the test?

Problem 3: You are running a simulation to estimate the power of a test where $H_0 : \mu = 100$. However, the oracle tells you the true population mean is actually $\mu = 105$. When you generate your random normal data in R using `rnorm()`, which mean do you use as the `mean` argument: 100 or 105?

Problem 4: Continuing Problem 3, after you generate the data, you need to calculate the test statistic $T = \frac{\bar{X} - \text{something}}{S/\sqrt{n}}$. What number must replace the "something" placeholder: 100 or 105?

Problem 5: You store 10,000 simulated p-values in a vector called `p_vals`. Write a single line of R code that calculates the estimated power of your test at a significance level of $\alpha = 0.05$.

Problem 6: True or False: When you are running a simulation study to estimate power, the histogram of your 10,000 simulated p-values will look perfectly flat, following a Uniform(0,1) distribution. Explain.

Problem 7: You are plotting a power curve where the x-axis is the sample size (n) and the y-axis is the estimated power. As the line moves from left to right (as n increases), should the line generally trend upwards or downwards? Why?

Problem 8: You are plotting a power curve where the x-axis is the true mean (μ). The null hypothesis being tested is $H_0 : \mu = 50$, and you are using $\alpha = 0.05$. When the x-axis is exactly at 50, what should the y-axis (power) value be?

Problem 9: In R, write the exact expression to calculate the rejection cutoff $t_{1-\alpha/2,n-1}$ for a two-sided test with $n = 40$ and $\alpha = 0.01$.

Problem 10: A classmate writes a simulation script to estimate power but accidentally generates the simulated data using the null mean μ_0 instead of the true mean μ. When they calculate the proportion of rejections at $\alpha = 0.05$, what number should they expect to see instead of the true power?

Practice Worksheet 12: Calculating Needed Replications

Core Concepts Review

The Big Picture: Why do we need this?

When we run a Monte Carlo simulation to estimate a probability (like Power or Type I Error), we are essentially guessing a hidden truth based on a random sample. If we run 100 replications, our guess might be pretty far off. If we run 1,000,000 replications, our guess will be incredibly accurate.

But computer time is valuable. We want to know: *Exactly how many replications (reps) do we need to guarantee our estimate is within a specific Margin of Error (m)?*

1. The Math Behind the Simulation

When we estimate Power or Type I Error, we are estimating a probability, which we will call $E(Y)$. In each replication of our simulation, our code either rejects the null hypothesis ($Y = 1$) or fails to reject it ($Y = 0$). This means Y is a Bernoulli random variable.

- **True Mean:** The true probability we want to find is $E(Y)$.
- **True Variance:** Because it is a Bernoulli variable, the variance is $E(Y)(1 - E(Y))$.

Thanks to the Central Limit Theorem, the average of all our simulated replications ($\bar{Y}$) follows a Normal distribution. The $100(1 - \alpha)\%$ confidence interval for our simulated estimate is:

$$\bar{Y} \pm z_{1-\alpha/2} \frac{\sqrt{E(Y)(1 - E(Y))}}{\sqrt{reps}}$$

2. Solving for *reps*

The right side of the plus/minus sign is our Margin of Error (m). If we set up the equation $m = z_{1-\alpha/2} \frac{\sqrt{E(Y)(1-E(Y))}}{\sqrt{reps}}$ and do some algebra to solve for $reps$, we get:

$$reps \geq \left(\frac{z_{1-\alpha/2} \sqrt{E(Y)(1 - E(Y))}}{m} \right)^2$$

3. The "Worst-Case Scenario" Trick

To solve the formula above, we need to plug in $E(Y)$. But wait—$E(Y)$ is the exact thing we are running the simulation to find! How can we plug it in if we don't know it?

- **The Conservative Approach:** The mathematical function $x(1-x)$ hits its absolute maximum value when $x = 0.5$. Therefore, the "worst-case" variance is $0.5 \times 0.5 = 0.25$. The square root of 0.25 is 0.5.
- **The Simplified Formula:** If we assume the worst-case scenario (that $E(Y) = 0.5$), the formula simplifies beautifully to guarantee our margin of error m:

$$reps \geq \left(\frac{z_{1-\alpha/2} \times 0.5}{m} \right)^2$$

- **The "Educated Guess" Approach:** If you know your test is good and the power $E(Y)$ is definitely greater than 0.80, the worst-case variance in that specific range happens at 0.80. You can plug in $\sqrt{0.8 \times 0.2}$ instead of 0.5 to save computer time and run fewer replications.

Practice Problems

Problem 1: When running a simulation to estimate a Type I error probability, the result of each replication is either 1 (Reject H_0) or 0 (Fail to Reject H_0). What specific named probability distribution does a single replication follow?

Problem 2: Using your answer from Problem 1, if the true power of a test is $E(Y)$, write the exact mathematical expression for the variance of a single replication.

Problem 3: You want to estimate the power of a test with a margin of error of at most $m = 0.02$ at a 95% confidence level ($z \approx 1.96$). Assuming you know absolutely nothing about what the true power might be, calculate the exact number of replications you need.

Problem 4: You are asked to estimate a Type I error probability with a margin of error of at most 0.005 at the 99% confidence level. Write the exact R code using the 'qnorm()' function to calculate the conservative worst-case number of replications required.

Problem 5: In R, if your calculation for *reps* results in 41234.12, what built-in R function should you wrap around your math to ensure you run a valid, whole number of replications that meets your requirement?

Problem 6: Suppose you want to estimate the power of a test, and previous studies guarantee the power is somewhere between 0.85 and 1.0. To calculate the needed replications, what specific numbers should you plug in for $E(Y)$ and $(1 - E(Y))$ to be as conservative as possible given this new information?

Problem 7: Use your numbers from Problem 6 to calculate the required *reps* if you want a margin of error of $m = 0.01$ at a 95% confidence level ($z \approx 1.96$).

Problem 8: Compare the conservative "worst-case" formula (using 0.5) to a situation where you know the true probability is exactly 0.5. Are they mathematically different, or do they yield the exact same number of required replications?

Problem 9: Look at the denominator of the *reps* formula: m. If your boss tells you to cut your margin of error strictly in half (e.g., from 0.02 to 0.01), by what exact factor will your required number of replications multiply? *Hint: Look at the exponent outside the parentheses.*

Problem 10: A classmate writes a simulation using $reps = 10,000$ to estimate a probability. They want to work backward to find out what their worst-case margin of error (m) is at the 95% confidence level ($z = 1.96$). Set up the algebra equation they need to solve to find m.

Practice Worksheet 13: Conditional Distributions & Intro to Regression

Core Concepts Review

The Big Picture: Moving Beyond the One-Sample Model

Up until now, we have studied the "One-Sample Model." This means we looked at a single random variable (like the temperature in Minneapolis) and tried to guess its mean (μ).

Regression takes a huge step forward. Instead of looking at a single variable in a vacuum, we want to see how one variable *depends* on another.

- **The Predictor (Input / x):** The variable we use to make our guess (e.g., the amount of fertilizer we apply). This is usually a fixed number chosen by the experimenter, not a random variable.
- **The Response (Output / Y):** The random variable we are trying to measure or predict (e.g., the resulting corn yield).

1. Conditional Distributions: $(Y|x)$

The notation $(Y|x)$ is read as "Y given x." It represents the distribution of the response Y *only for* a specific, targeted value of x.

For example, if we apply exactly 50 lbs of fertilizer ($x = 50$), the corn yield isn't going to be exactly the same every single time. It will fluctuate. That fluctuation is the conditional distribution $(Y|50)$.

- $E(Y|x)$ is the expected average yield *specifically when* we apply x fertilizer.
- $sd(Y|x)$ is the standard deviation (fluctuation) *specifically when* we apply x fertilizer.

2. The Regression Model

If we only tested 5 specific amounts of fertilizer, we could just calculate 5 separate sample means. But what if we want to predict the corn yield for an amount of fertilizer we *didn't* test?

To do this, we build a mathematical model that links x and Y together:

$$Y_i = \mu_Y(x_i) + \sigma_Y(x_i)Z_i$$

- $\mu_Y(x_i)$ is the **Population Mean Function**. Instead of a single flat mean, the mean is now a curve or a line based on x. (e.g., $\mu_Y(x) = \beta_1 + \beta_2 x$).
- $\sigma_Y(x_i)$ is the **Population Standard Deviation Function**. We frequently assume this is just a constant c, meaning the data fluctuates the same amount no matter what x is.
- Z_i is a random error term with a mean of 0 and a variance of 1.

By using a simple function for the mean (like a straight line with just a slope and intercept), we can accurately estimate it without needing thousands of data points for every single possible value of x.

Practice Problems

Problem 1: In your own words, what is the fundamental difference between the "one-sample model" and a "regression model"?

Problem 2: You want to study how the number of hours spent studying impacts a student's final exam score. Identify the predictor variable and the response variable.

Problem 3: Explain what the conditional expected value $E(Y|x = 10)$ represents in the context of Problem 2.

Problem 4: In the general regression model $Y_i = \mu_Y(x_i) + \sigma_Y(x_i)Z_i$, explain why the predictor values $x_1, \ldots, x_n$ are generally written with lowercase letters rather than uppercase letters.

Problem 5: A researcher tests 6 different temperatures and measures the resulting chemical yield 5 times at each temperature. Why is it generally a bad idea to just connect the 6 sample means with a line to form a predictive model, rather than using a continuous function like $\mu_Y(x) = \beta_1 + \beta_2 x$?

Problem 6: Suppose you assume the population mean function is a straight line: $\mu_Y(x) = \beta_1 + \beta_2 x$. What do the parameters β_1 and β_2 conceptually represent?

Problem 7: After analyzing your data, R gives you the estimated population mean function: $\hat{y}(x) = 15.2 + 3.4x$. Predict the expected response if you were to set your predictor variable to $x = 5$.

Problem 8: Many regression models assume a "constant population standard deviation function," meaning $\sigma_Y(x) = \sigma$. Explain what this assumption means visually if you were to look at a scatterplot of your data.

Problem 9: Suppose you are looking at a time plot of weather data over 50 years, and there is no upward or downward trend, and the variability stays exactly the same. Does this data require a regression model to find its mean, or is the one-sample model sufficient? Why?

Problem 10: In R, if you have a dataset where the response is stored in `corn$yield` and the predictor is stored in `corn$nitrogen`, what built-in R function can you use to instantly calculate the separate sample means for every unique level of nitrogen tested? *Hint: It evaluates a function "by" groups.*

Practice Worksheet: R Code Fill-in-the-Blank

Simulation and Distributions

Problem 1: Suppose you want to generate a realization of $Y \sim Binomial(n = 10, \theta = 0.2)$. Fill in the missing R code.

```
X <- runif(10, min = -1, max = 1)
Y <- sum( X <= ????? )
```

Problem 2: Again, generate $Y \sim Binomial(n = 10, \theta = 0.2)$ using a Normal distribution. Fill in the missing R code.

```
X <- 1 * ( rnorm(10) >= ????? )
Y <- sum(X)
```

Problem 3: Generate a single realization of $Y \sim Binomial(n = 10, \theta = 0.2)$ using a matrix of Bernoulli trials.

```
X <- matrix( ????? , nrow = 20, ncol = 10)
Z <- apply(X, 1, sum)
Y <- Z[1]
```

Quantiles and QQ-Plots

Problem 4: You have 10,000 realizations of a test statistic stored in `t_vals`. You want to create a QQ-plot to see if they follow a t-distribution with 19 degrees of freedom.

```
probs <- ?????(10000)
theoretical_quantiles <- ?????(probs, df = 19)
data_quantiles <- quantile(t_vals, probs)
plot(theoretical_quantiles, data_quantiles)
abline(0, 1)
```

Confidence Intervals and Power

Problem 5: You want to construct a 95% confidence interval for the mean using the t-distribution. Your data is in `x.list`.

```
n <- length(x.list)
xbar <- mean(x.list)
moe <- ?????(0.975, df = n - 1) * sd(x.list) / sqrt(n)
```

Problem 6: You are estimating the power of a test via simulation. You have 5,000 p-values stored in `pvals`, and your significance level is $\alpha = 0.05$. Write the single command to estimate power.

```
est_power <- ?????(pvals < 0.05)
```

Problem 7: Calculate the worst-case number of replications needed to estimate a Type I error probability with a margin of error of at most 0.005 at the 99% confidence level.

```
reps <- ceiling(( ?????(0.995) * 0.5 / 0.005 )^2)
```

Practice Worksheet: Conceptual True/False

True or False Questions

Directions: Determine whether each statement is TRUE or FALSE. Provide a brief explanation for your reasoning.

Problem 1: Suppose $Y \sim Exp(\mu)$. Then $P(-2 \leq Y \leq 2) > P(0 \leq Y \leq 2)$.

Problem 2: Suppose $Y \sim Exp(\mu)$. Then $P(0 \leq Y \leq 2) = P(0 \leq Y < 2)$.

Problem 3: For any continuous random variable Y, if interval A is wider than interval B, then $P(Y \in A) \geq P(Y \in B)$.

Problem 4: If a 95% confidence interval for a population mean μ covers zero, this means that with 95% confidence, zero is a plausible value of the mean based on our sample.

Problem 5: Suppose we calculate a 95% confidence interval for a population mean and get $[10.2, 15.4]$. The probability that the true mean μ lies in this specific interval is exactly 0.95.

Problem 6: Suppose the data $X_1, \ldots, X_{10}$ are iid from a Uniform distribution. The random confidence interval $\bar{X} \pm t_{1-\alpha/2,9} \frac{S}{\sqrt{10}}$ will cover the true mean with probability exactly $1 - \alpha$.

Problem 7: Relative to when n is small, a 95% confidence interval should have a coverage probability strictly greater than 0.95 when n is sufficiently large.

Problem 8: The observed p-value of a hypothesis test represents the probability that the null hypothesis (H_0) is true.

Problem 9: When the null hypothesis is true and the data comes from a Normal distribution, the random p-value of a one-sample t-test perfectly follows a $Unif(0, 1)$ distribution.

Problem 10: Let S^2 be the sample variance of n independent observations from a distribution with true variance σ^2. The expected value of S^2 is exactly equal to σ^2.

Practice Midterm Exam (Topics 1-5)

STAT3301 Regression and Statistical Computing

Directions: This exam covers Probability Foundations, Named Distributions, Estimators, the Inversion Method, and Monte Carlo Simulations. Show all work for full credit.

---

1. (6 pts) The Inversion Method and Percentiles

Suppose that the random variable X has cumulative distribution function (cdf) given by:

$$F_X(t) = P(X \leq t) = \begin{cases} 0 & t < 0 \\ 1 - (1-t)^3 & 0 \leq t \leq 1 \\ 1 & t > 1 \end{cases}$$

(a) (4 pts) Suppose $U \sim Uniform(0, 1)$. Using the inversion method, derive an exact mathematical expression for a random variable—depending only on U—that has the exact same distribution as X.

(b) (2 pts) Calculate the exact median (the 50th percentile) of the random variable X.

2. (6 pts) Expected Value, Variance, and Probability Properties

Let $Y \sim Uniform(-2, 4)$.

(a) (3 pts) Using the properties of expected value and variance, compute the exact value of $E(Y^2)$.

(b) (3 pts) Answer the following TRUE or FALSE questions. Give a brief 1-sentence explanation for each.

- **TRUE / FALSE:** $P(0 \leq Y \leq 2) = P(0 < Y < 2)$.
  Explanation:
- **TRUE / FALSE:** If we generate data using `y <- runif(1000, -2, 4)`, the result of `mean(y)` will be exactly 1.0.
  Explanation:
- **TRUE / FALSE:** The variance of Y, $var(Y)$, is a random variable.
  Explanation:

3. (4 pts) Estimators vs. Estimates

Suppose a researcher measures the weights of 5 randomly selected rocks, and records the following values (in grams): $x_1 = 10, x_2 = 12, x_3 = 15, x_4 = 11, x_5 = 12$.

(a) (2 pts) The researcher calculates $s^2 = \frac{1}{4}\sum_{i=1}^{5}(x_i - 12)^2 = 3.5$. Is 3.5 an estimator or an estimate? Is 3.5 a random variable?

(b) (2 pts) Let $X_1, \ldots, X_5$ be the unobserved random variables representing the rock weights, and assume they come from a population with a true variance of $\sigma^2 = 4$. What is the expected value of the sample variance estimator, $E(S^2)$?

4. (5 pts) R Coding and Simulation (Fill-in-the-Blank)

Fill in the missing R code labeled with `?????` to correctly complete each task.

(a) (2 pts) You want to generate a single realization of $Y \sim N(\mu = 10, \sigma^2 = 25)$ using the Box-Muller method.

```
U1 <- runif(1)
U2 <- runif(1)
Z <- sqrt(-2 * log(U1)) * cos(2 * pi * U2)
Y <- ?????
```

(b) (2 pts) You want to run a Monte Carlo simulation to estimate the probability that a $Unif(0, 10)$ random variable is strictly greater than 7.

```
set.seed(3301)
sim_data <- runif(10000, min = 0, max = 10)
estimated_prob <- ?????
```

(c) (1 pt) You have a matrix `X_mat` with 10,000 rows and 15 columns representing a simulation study. You want to calculate the sample mean of each row and store it in `means_list`.

```
means_list <- apply(X_mat, ?????, mean)
```

5. (4 pts) Conceptual Monte Carlo

(a) (2 pts) In a Monte Carlo simulation, we often rely on the Law of Large Numbers (LLN). Briefly explain what the LLN guarantees as the number of replications (`reps`) approaches infinity.

(b) (2 pts) Why is it necessary to use the `set.seed()` function before running a simulation study that you plan to submit for grading?

Practice Midterm Exam (Topics 6-9)

STAT3301 Regression and Statistical Computing

Directions: This exam covers the Central Limit Theorem, QQ-Plots, Confidence Intervals for the Mean, and Confidence Intervals for Bernoulli Proportions. Show all work for full credit.

---

1. (5 pts) Central Limit Theorem and QQ-Plots

Suppose $X_1, \ldots, X_{60}$ are independent and identically distributed random variables from an Exponential distribution with a true mean of $\mu = 10$ and true variance of $\sigma^2 = 100$. Let $\bar{X}$ be the sample mean.

(a) (2 pts) According to the Central Limit Theorem, what is the approximate distribution of $\bar{X}$? Provide the name of the distribution and its exact parameters.

(b) (3 pts) Suppose you create a Normal QQ-plot for a dataset of simulated test statistics, plotting the theoretical Normal quantiles on the x-axis and the data quantiles on the y-axis. You notice the points on the far right of the plot dip noticeably *below* the $y = x$ reference line. What does this indicate about the right tail of your data's distribution compared to a true Normal distribution?

2. (6 pts) Confidence Intervals for the Mean

A researcher measures the lifespans of 25 batteries. The sample mean is $\bar{x} = 40$ hours, and the sample standard deviation is $s = 5$ hours.

(a) (3 pts) Write the exact, calculator-ready mathematical expression to compute the 95% confidence interval for the true mean lifespan μ. (Use $t_{p,\nu}$ notation for any necessary distribution quantiles, being specific about the parameter values).

(b) (3 pts) Answer the following TRUE or FALSE question and give a brief 1-sentence explanation.
TRUE / FALSE: The true mean μ is a random variable that falls within your calculated interval from part (a) with exactly a 95% probability.
Explanation:

3. (5 pts) Confidence Intervals for Bernoulli Success Probability

Suppose you observe $n = 100$ independent Bernoulli trials and record 36 successes (so $\bar{x} = 0.36$). You want to estimate the true success probability, θ.

(a) (3 pts) Using $z_{0.975} \approx 1.96$, write the exact, calculator-ready mathematical expression for the 95% **Simple Conservative** confidence interval for θ.

(b) (2 pts) Explain mathematically why the Simple Conservative interval will *always* be at least as wide as the Wald interval for the exact same dataset.

4. (5 pts) R Coding (Fill-in-the-Blank)

Fill in the missing R code labeled with `?????` to correctly complete each task.

(a) (2 pts) You have a vector of 500 simulated sample means stored in `ybar_list`. You want to calculate the coordinates for a Normal QQ-plot.

```
probs <- ?????(500)
theoretical_quantiles <- ?????(probs, mean = mean(ybar_list),
                               sd = sd(ybar_list))
```

(b) (1 pt) You have a binary vector of data `x_vals` of length 50. Write the missing R code to compute the Score interval using a built-in R function.

```
score_ci <- ?????(x = sum(x_vals), n = 50, correct = FALSE)$conf.int
```

(c) (2 pts) You want to find the exact t-distribution multiplier $t_{0.995,14}$ for a 99% confidence interval based on a sample size of 15.

```
t_mult <- ?????(0.995, df = ?????)
```

5. (4 pts) Conceptual True/False

Answer the following TRUE or FALSE questions. Provide a brief 1-sentence explanation for your choice.

(a) (2 pts) **TRUE / FALSE:** In simulation studies, the Wald interval for a Bernoulli success probability θ frequently "under-covers", meaning its actual coverage probability is strictly lower than the requested $1 - \alpha$ target.
Explanation:

(b) (2 pts) **TRUE / FALSE:** If you collect a much larger sample of data (increasing your sample size n from 20 to 200) while keeping the sample standard deviation s exactly the same, the overall width of your one-sample t-interval will increase.
Explanation:

Practice Midterm Exam (Topics 10-13)

STAT3301 Regression and Statistical Computing

Directions: This exam covers Type I/II Errors, p-values, Power, Simulation Replications, and the Introduction to Regression. Show all work for full credit.

---

1. (5 pts) Hypothesis Testing and Errors

Suppose you perform a one-sample t-test for $H_0 : \mu = 50$ vs $H_a : \mu \neq 50$ at a significance level of $\alpha = 0.05$.

(a) (2 pts) You observe a p-value of 0.02 and therefore decide to reject H_0. If an oracle later tells you the true population mean was actually exactly 50 all along, what specific type of error did you commit?

(b) (3 pts) Answer the following TRUE or FALSE question and give a brief 1-sentence explanation.
TRUE / FALSE: When the null hypothesis H_0 is completely true, the distribution of the random, yet-to-be-observed p-value is exactly $Uniform(0, 1)$.
Explanation:

2. (5 pts) Power and Simulation

(a) (2 pts) Suppose you are plotting a power curve for a one-sample t-test where the x-axis is the sample size (n) and the y-axis is the estimated power. As n increases, what should happen to the power curve? Why?

(b) (3 pts) You run a Monte Carlo simulation to estimate the power of a test at a significance level of $\alpha = 0.01$. You have already generated 10,000 p-values under the alternative hypothesis and stored them in a vector called `pvals`. Fill in the missing R code labeled with `?????` to compute the simulated estimate of power.

```
est_power <- ?????(pvals < ?????)
```

3. (5 pts) Calculating Needed Replications

You want to estimate the true power of a test via simulation. You require a margin of error of at most $m = 0.02$ at the 95% confidence level (where $z_{0.975} \approx 1.96$).

(a) (3 pts) Assuming you have absolutely no prior knowledge about what the true power might be, write the exact, calculator-ready mathematical expression to calculate the conservative, worst-case number of replications ($reps$) required.

(b) (2 pts) Fill in the missing R code labeled with `?????` to compute this worst-case value and ensure it evaluates to a valid, whole number of replications.

```
reps <- ?????((qnorm(0.975) * ????? / 0.02)^2)
```

4. (5 pts) Introduction to Regression

A researcher wants to model crop yield (Y) based on the amount of fertilizer applied (x) using the regression model $Y_i = \mu_Y(x_i) + \sigma_Y(x_i)Z_i$.

(a) (2 pts) In your own words, explain exactly what the conditional expected value $E(Y|x = 20)$ represents in the real-world context of this problem.

(b) (3 pts) After collecting data and fitting a linear model, R provides the estimated sample mean function: $\hat{y}(x) = 45 + 2.5x$. What is your predicted crop yield if you were to apply $x = 10$ units of fertilizer?

5. (5 pts) Conceptual True/False

Answer the following TRUE or FALSE questions. Provide a brief 1-sentence explanation for your choice.

(a) (2.5 pts) **TRUE / FALSE:** If a statistical test has a Type II error probability of $\beta = 0.20$, then the power of that test is exactly 0.80.
Explanation:

(b) (2.5 pts) **TRUE / FALSE:** The observed p-value of a hypothesis test represents the probability that the null hypothesis (H_0) is true.
Explanation: