Sampling Distribution of Pearsons $\hat{r}$

1) Shape of the sampling Distribution of $\hat{r}$

- Pearson's $\hat{r}$, which measures the strength and direction of a linear relationship between two variables, does not follow a normal distribution

Key observation The distribution of r is negatively skewed. The skew The skew occurs because r can never be greather than 1.0, limiting the distributions range in the positive direction. As the population correlation (ρ) increases, the skew becomes more pronounced.

Examples

- for a population correlation of $\rho = 0.60$, the distribution is moderately skewed
- for a population correlation of $\rho = 0.90$, the distribution is sharply skewed, with a short positive tail and a long negative tail.

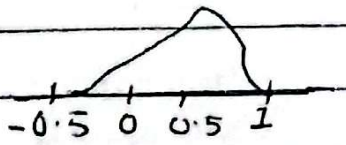

Figure 1

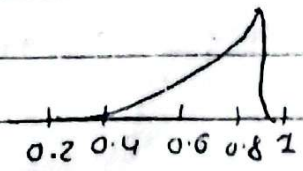

Figure 2

Figure 1: Distribution of r for $N=12$ and $\rho=0.60$ is negtively skewed
Figure 2: Distribution of r for $N=12$ and $\rho=0.90$ is more pronouncedly skewed with a short positive tail and long negative tail.

2) Transforming r to z'

- Since the sampling distribution of r is not normal, a transformation is necessary to convert r into a variable that follows a normal distribution. This transformed variable is denoted as z'

Transformation formula:

$$z' = 0.5 \times \ln\left(\frac{1+r}{1-r}\right)$$

where

r is the sample correlation

ln represents the natural logarithm

Why use z'

- z' follows a normal distribution and has a standard error of

$$\text{Standard Error of } z' = \frac{1}{\sqrt{N-3}}$$

where N is the number of pairs of scores in the sample.

3) Computing the Standard Error of z'

To compute standard error of z' use the formula

$$\text{Standard Error of } z' = \frac{1}{\sqrt{N-3}}$$

Example

for N=12 (sample of students) the standard error of z' is

$$= 0.333$$

4) Calculating the probability of an r above a specific value

Problem: Suppose we have a population correlation of $\rho = 0.6$, and we want to determine the probability that in a random sample of 12 students, the sample correlation r will be 0.75 or higher

1) Steps:
Transform both values to z'
for $p = 0.60$ the transformation is 0.643
for $r = 0.75$, the transformation is 0.973

2) Compute the standard error of z' for $N = 12$
standard error of $z' = 0.333$

3) Determine the z-score for $r = 0.75$ (ie $z' = 0.973$) = 0.84

4) Find the probability using a z-table or a calculator, the area above a z-score of 0.84 is 0.20
Therefore, the probability of obtaining a sample correlation $r \geq 0.75$ is 0.20

Sampling distribution of p
1 Computing the Mean and Standard Deviation of the Sampling Distribution of P
The sampling distribution of p is the distribution that would result if we repeatedly took random samples of size N and recorded the sample proportion (p) of a particular outcome.
Example: In an election, 60% ($\pi = 0.60$) of voters prefer candinte A. A random sample of 10 voters may not always perfectly match this proportion.

Mean of the Sampling distribution of p: since p represents a sample proportion, the expected mean of its distribution is simply the population proportion (π):
$\mu_p = \pi$

For this example $\mu_{\hat{p}} = \pi$

Discrete vs Continuous Nature

- The sampling distribution of $\hat{p}$ is discrete when N is small
- For example, with $N = 10$, $\hat{p}$ can take on values such as 0.50 (5 out of 10 voters) or 0.60 (6 out of 10 voters) but not 0.55

- Approximation to the Normal Distribution

The sampling distribution of $\hat{p}$ becomes approximately normal if N is large enough and π is not too close to 0 or 1

A good rule of thumb:

The normal approximation is valid if:

$$N\pi \geq 10 \text{ and } N(1-\pi) \geq 10$$

Basic Sample Statistics and Parameters

Introduction to Estimation

Key Definitions

1) Statistic - A numerical value calculated from a sample (eg sample mean, sample proportion).
2) Parameter - A numerical value that describes a parameter population (eg population mean, population proportion)
3) Point Estimate - A single value used to estimate a population parameter
4) Interval Estimate - A range of values likely to contain the population parameter.
5) Margin Error: The range within which the true population parameter is expected to fall, based on a confidence level

Point Estimates

- A point estimate is a single number used to estimate a population parameter
- Example: A poll surveys 200 people and 106 support building a new sports stadium
  - Sample proportion (point estimate) = 106/200 = 0.53
  - This means we estimate that 53% of the population supports the proposition

Confidence Intervals (Interval Estimates)

A confidence interval provides a range of values that is likely to contain the population parameter

Example: A 95% confidence interval for the stadium poll is $0.46 < p < 0.60$

- This means we are 95% confident that the proportion of the population supporting the proposition is between 46% and 60%.
- In media reports, this is often stated as "53% favors the proposition with a margin of error of 7%"

Degrees of Freedom

Key Definitions

1) Degrees of Freedom (df) - The number of independent pieces of information used to estimate a parameter.
2) Variance Estimation - The process of calculating variance based on sample data.
3) Independence of Deviations - Whether individual values contribute independently to an estimate.

4) General Formula for Degrees of Freedom - df = Number values - Number of estimated parameters.

Understanding Degrees of freedom:

- Estimates based on larger sample sizes are more accurate because they use more independent pieces of information

Example: Estimating Martian height variance.

- If the population mean is known:
  - sample 1 Martian (height = 8)
  - variance estimate: $(8-6)^2 = 4 \rightarrow 1$ degree of freedom (df = 1).
  - sample 2 Martians (heights = 8 and 5)
  - variance estimates $(8-6)^2 = 4$, $(5-6)^2 = 1$
  - Average variance estimate $(4+1)/2 = 2.5$
    $\rightarrow 2$ degrees of freedom

General Formula for Degrees of Freedom

$df = N - 1$, where N is the number of observations

Example: If 12 Martians were sampled, the degrees of freedom would be $12 - 1 = 11$

Variance formula

- The formula for estimating variance in a sample

$$s^2 = \frac{\sum (x - M)^2}{N - 1}$$

- The denominator $(N-1)$ represents the degrees of freedom.

Characteristics of Estimators

Key Concepts

1) Bias - When an estimator consistently overestimates or underestimates the true parameter
2) Sampling Variability - The extent to which an estimate changes from sample to sample.
3) Expected value - The long term average of a statistic which should ideally equal the population parameter.
4) Relative Efficiency - A measure comparing the variability of two different estimators

Understanding Bias

Example: Two Bathroom Scales

- Scale 1 (Biased but Precise): consistently overestimates weight by 1 pound but varies little.
- Scale 2 (Unbiased but variable): sometimes overestimates, sometimes underestimates but averages to the correct weight.

Formal Definition

- A statistic is biased if it expected value (long-term average) does not equal the parameter it estimates.
- The sample mean is unbiased because its expected value equals the population mean (μ)
- The sample variance (s^2) would be biased if divided by N instead of $N-1$. Using $N-1$ corrects this bias, making it an estimator of population variance (σ^2)