

Basics

	Sample statistic	Population Parameter
Mean	\bar{X}	μ
Variance	S^2	σ^2
Correlation	r	ρ
Noise	e_i	ϵ_i
	Guess	True, but unknown

The Mean, Variance and StdDev are subject to outliers.

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\mu = E(W) = E(a + bX) = a + bE(X)$$

$$\sigma^2 = Var(X) = E[X^2] - \mu_x^2$$

$$S^2 = \sum_{all\ i} \frac{(X_i - \bar{X})^2}{n - 1}$$

$$\sigma = StdDev(X) = \sqrt{\sigma^2}$$

$$Var(X) = E[(X - \mu)^2] = E(X^2) - \mu^2$$

$$= \sum_{all\ x} (x - \mu_x)^2 P(X = x)$$

$$= E(X^2) - E(X)^2 = E(X - E(X))^2$$

$$Var(X + c) = Var(X)$$

$$Var(cX) = c^2 Var(X)$$

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

If X,Y are dependent

$$Var(X + Y) \neq Var(X) + Var(Y)$$

$$Var(X + Y) = Var(X - Y)$$

$$Var((a + bX) + (c + dY)) = b^2 Var(X) + d^2 Var(Y) + 2bd Cov(X, Y)$$

Quartiles split the data into 4 equal groups by number of values, or 25% percentiles. Q2 is the median.

Covariance and Correlation

Covariance gives direction.

Correlation gives direction and strength.

Both are *linear*.

$$Cov(X, Y) = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})$$

$$Cor = \frac{Cov}{\sigma_x \cdot \sigma_y}$$

Probabilities

$0 \leq$ All probabilities ≤ 1

Mutually exclusive and Exhaustive

$$P(\bar{A}) = 1 - P(A)$$

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

Joint vs. Marginal probabilities

Independent if $P(A|B) = P(A)$

For Independent only: $P(E \text{ and } F) = P(E) \cdot P(F)$

	B	\bar{B}
A	$P(A \text{ and } B) = P(B)P(A B)$	$P(A \text{ and } \bar{B}) = P(\bar{B})P(A \bar{B})$
\bar{A}	$P(\bar{A} \text{ and } B) = P(\bar{A})P(B \bar{A})$	$P(\bar{A} \text{ and } \bar{B}) = P(\bar{A})P(\bar{B} \bar{A})$
all success	p^n	
all failure	$(1 - p)^n$	
at least one failure	$1 - p^n$	
at least one success	$1 - (1 - p)^n$	

Random Variables

$$P(X \leq x) \rightarrow CDF$$

$$E(cX) = c \cdot E(X)$$

$$E(X + c) = E(X) + c$$

$$E(X + Y) = E(X) + E(Y)$$

Bias of an Estimator

In practice n has to be relatively much larger like > 100 .

Guesses should be *unbiased* and have *minimum variance*.

MVUE (Minimum Variance, Unbiased Estimates).

$$bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Unbiased if $bias = 0$ (expected value equals true, not a particular value of \bar{x}).

For samples, we divide by $n - 1$ instead of n to make an unbiased estimator. The guess would otherwise be too low.

$$E(X) = \sum_{all\ x} xP(X = x)$$

$$E(\bar{X}) = \sum_{i=1}^{\infty} x_i p_i = \mu$$

$$E(S^2) = \sigma^2$$

$$E[X + c] = E[X] + c$$

$$E[X + Y] = E[X] + E[Y]$$

$$E[aX] = aE[X]$$

$$E[aX + bY + c] = aE[X] + bE[Y] + c$$

$$E((a + bX) + (c + dY)) = a + bE(X) + c + dE(Y)$$

Example: Roulette has a \$1 bet with a \$35 payoff for $\frac{1}{38}$ odds.

$$E[\text{gain from a \$1 bet}] = -\$1 \cdot \frac{37}{38} + \$35 \cdot \frac{1}{38} = -\$0.0526$$

All Things χ^2

$$\chi_1^2 = Z^2$$

$$\chi_n^2 = \sum_{i=1}^n Z^2$$

$$t_n = \frac{Z}{\sqrt{\chi_n^2/n}}$$

$$F_{n,m} = \frac{\chi_n^2/n}{\chi_m^2/m}$$

$$\chi_{n+m}^2 = \chi_n^2 + \chi_m^2$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{df=n-1}^2$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{x})^2 \sim \frac{\sigma^2}{n-1} \chi_{df=n-1}^2$$

The t distribution

The t distribution has a slightly wider spread in the tails than Standard Normal. Also known as student-t distribution. This happens when you don't know the real σ and are forced to use S , which is less reliable.

$$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

allows us to Z-score

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

but we don't know σ so in practice we use

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

(which has two random variables (\bar{X} and S)). Now, take that previous definition of T and add σ/\sqrt{n} to the top and bottom:

$$\begin{aligned} T &= \frac{\bar{X} - \mu}{S/\sqrt{n}} \\ &= \frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{(S/\sqrt{n})/(\sigma/\sqrt{n})} \\ &= \frac{Z}{\sqrt{S^2/\sigma^2}} \\ &= \frac{Z}{\sqrt{\frac{(n-1)S^2}{(n-1)\sigma^2}}} \\ &= \frac{Z}{\sqrt{\chi^2/df}}, \text{ since } \chi^2 = \frac{(n-1)S^2}{\sigma^2} \end{aligned}$$

Hypothesis Test - General

1. Formulate hypotheses
2. Calculate test statistic
3. Calculate p -value based on reference distribution of test statistic $|H_0$. **p value is NOT the probability that the null hypothesis is true.** It is probability of seeing our data on the null hypothesis.
4. Determine conclusion and scope of inference

Note: t values are measures of *distance*. p values are measure of *probability*.

Hypotheses	Decision Rule
$H_0 : \mu = \mu_0$	
$H_a : \mu \neq \mu_0$	If $ t_{stat} > 1.96$, reject H_0

$H_0 : \mu = \mu_0$	
$H_a : \mu < \mu_0$	If $t_{stat} < -1.64$, reject H_0

$H_0 : \mu = \mu_0$	
$H_a : \mu > \mu_0$	If $t_{stat} > 1.64$, reject H_0

Scope of Inference: Interval Validity: low when 1) unaccounted confounding factors, 2) ignored missing data 3) noncompliance 4) unverified assumptions 5) suboptimal method of analysis. Applicable when allocation of units to groups is random. **External Validity:** high when can be generalized to population. Applicable when selection of units is random.

Assumptions:

1. Observations are **independent**. This is true if the sample is randomly selected, but false if there is bias in the sampling.
2. **Normal Distributions** of the observations. This happens when the sample is large enough (via the Central Limit Theorem) or it is known that the population is normally distributed.

If these assumptions are not correct, none of the t -tests will work correctly.

Hypothesis Test - Mean

Calculation by hand using a t test:

$$t_{stat} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Types of Errors

Type I the null hypothesis is rejected when it is true

Type II the null hypothesis is accepted when it is false

α is *level of significance* - probability of making a Type I error. The greater the cost of an error, the smaller α should be. β is the probability of making a Type II error. There is an inverse relation between Type I and II errors. Reducing one increases the other. The only way to reduce both is to increase n the sample size.

Confidence Intervals

CI is Estimate \pm margin of error or Estimate $\pm (z \text{ value}) \times$ (SD of estimate).

$$\mu_0 = \bar{X} \pm t \cdot s/\sqrt{n}$$

If the interval spans 0, then one is not significantly bigger than the other (or cannot be determined). As long as $n > 30$, it doesn't matter if the sample size is different between the random variables.

To find the 95% Confidence Interval for the **true mean μ** :

$$\bar{x} \pm t^*(s/\sqrt{n})$$

t^* has to be looked up

Comparing Two Proportions

$$Var(\hat{p}_1 - \hat{p}_2) = \frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}$$

The 95% confidence interval for $p_1 - p_2$ is:

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

Decision Rules for Testing Two Proportions:

$$T = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1 - \hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}, \text{ where } \hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

\hat{p} is called the *pooled proportion*.

Comparing Two Means

Requirements:

1. σ_1 and σ_2 are unknown. No assumption made about their equality.
2. The two samples are independent.
3. Both samples are simple random samples.
4. The two samples size are both large (ie. > 30) or both populations have normal distributions.

A confidence interval for $(\mu_1 - \mu_2)$ is

$$(\bar{x}_1 - \bar{x}_2) \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Comparing Two Normal Distributions

$$A \sim \mathcal{N}(\mu_A, \sigma_A^2)$$

$$B \sim \mathcal{N}(\mu_B, \sigma_B^2)$$

Find $P(A < B + c)$:

$$P(A - B - c < 0)$$

$$(A - B - c) \sim \mathcal{N}(\mu_A - \mu_B - c, \sigma_A^2 + \sigma_B^2 - 2Cov(A, B))$$

$$P(X = A - B - c > 0)$$

then Z-score

Matched Pairs

This when there are two samples that are **not** independent, e.g. Weight Watchers, Before / After or matched, shared characteristics.

Is the data matched or independent?

If we don't take into account the match, the results are wrong. To account for this, take the difference between $\bar{X}_1 - \bar{X}_2$ and then do a hypothesis test on the *difference*.

$$H_0 : \mu_D = 0$$

$$H_a : \mu_D > 0$$

Uniform Distribution

$$E(X) = \frac{(a + b)}{2}$$

$$Var(X) = \frac{(b - a)^2}{12}$$

y axis should be a fraction to make area = 1

Normal Distribution

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

$$X = \sigma Z + \mu$$

$$P(a \leq X \leq b) = P[(a - \mu) \leq (X - \mu) \leq (b - \mu)]$$

$$= P\left[\frac{(a - \mu)}{\sigma} \leq \frac{(X - \mu)}{\sigma} \leq \frac{(b - \mu)}{\sigma}\right]$$

$$= P\left[\frac{(a - \mu)}{\sigma} \leq Z \leq \frac{(b - \mu)}{\sigma}\right]$$

Binomial Distribution

- n independent trials
- binary result
- same probability of success
- total number of successes

$$X \sim Bin(n, p)$$

$$\binom{n}{x} = \frac{n!}{x!(n - x)!}$$

$$P(X = x) = \frac{n!}{x!(n - x)!} p^x q^{(n - x)}$$

$$\mu_x = E(X) = n \cdot p$$

$$\sigma^2 = Var(X) = n \cdot p \cdot q = n \cdot p \cdot (1 - p)$$

Shape of distribution depends on p, n .

Small p , left skewed. Large p , right skewed

Fisher's Randomization Test

Randomization Test is for experiments - for casual inference.
Permutation Tests are for observational studies - for generalization.

Fisher Randomization Test is a *distribution free* test for treatment effect in **randomized experiments**. No assumption as to the underlying distribution has to be made. Generally, is $\mu_1 = \mu_2$?

The only assumption is Additive Treatment Effect (δ):

$$Y_{c,i} = Y_{v,i} + \delta$$

$H_0 : \delta = 0$, ie zero treatment effect for all units. Each unit's outcome is the same regardless of the treatment assigned. So the distribution would be identical in both groups.

$H_a : \delta \neq 0$ non-zero treatment effect for ALL units.

Assumptions:

- Random assignment to groups
- Under H_0 , *independence, interchangability* of study units.

Test Statistic

Difference of outcomes of the two groups:

$$\hat{\delta} = \bar{Y}_c - \bar{Y}_v$$

could also use difference between the medians.

Randomization Distribution is the reference distribution of a test statistic in a randomization test, where variation is due to random assignment of the treatment.

Procedure: Y is values of malaria. X is if someone is vaccinated. Simulate a new X^* with the same values as X but *in a different order*, leaving the same Y . Repeat multiple times to simulate and build the histogram. This will show if the distribution is extreme.

The maximum number of possible simulations is the binomial distribution:

$$\binom{t}{s}$$

where t is the total number of study units and s is the number where the effect was shown.

Calculating p value

The p value is proportion of values above or below the observed test statistic.

$$\hat{\delta} = 0.5$$

$$p = P(\bar{Y}_c - \bar{Y}_v \geq \bar{y}_c - \bar{y}_v)$$

$$= P(\bar{Y}_c - \bar{Y}_v \geq 0.5) = 0.029$$

so there is sufficient evidence to reject the null.

Scope of inference: Internal Validity maybe not if the same hospital. External validity: possibly not, as these were volunteers.