BIOSTATISTICS - formulae

Define the following terms – Probability, Experiment, outcomes, Sample Space, Event

1. Probability

(i) Events

- Of two mutually exclusive events → **$p(A \cup B) = p(A) + p(B) = 1$**
- Of two independent events-- **$p(A \cap B) = p(A) * p(B)$**

(ii) Probability distribution of variables

1. Normal distribution - (Continuous variables)-- First standardize variable to standard normal deviate (Z-score), **then interpret Z-Score**

$$Z = \frac{Y - \mu}{\sigma}$$

2. Binomial Distribution - (Binary Variables)

$$p(Y = y) = \binom{n}{y} \pi^{y} (1 - \pi)^{n-y}$$

Where:

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

Where n (number of trials/observations) is large the distribution of the binomial variable and the proportion are approximately **normal when:**

$$n\pi \geq 5$$

and $n(1 - \pi) \geq 5$

The mean and variance of the binomial distribution can then be calculated as follows:

$$\mu = n\pi$$

$$\sigma 2 = n\pi(1 - \pi)$$

The equation for normal distribution can then be applied

3. Poisson Distribution- (Discrete/Rare occurrences)

$$p\,(Y = y) = \frac{\lambda^{y} e^{-\lambda}}{y!}$$

2. Statistical Inference

Conclusions about a population are made based on findings from sample of the population.

Measures of Statistical Inference:

- Confidence Interval (Commonly at 95% CI)
  CI of Means (use Z-score) ***Use t distribution table when sample size < 20***

$$CI = \bar{X} \pm 1.96 \times SE\,(\bar{X})$$

Standard Error of mean $= \frac{SD}{\sqrt{n}}$

CI of Proportions (use Z-score) ***Use t distribution table when sample size < 20***

$$CI = \bar{X} \pm 1.96 \times SE(\bar{p})$$

Standard Error of Proportion $= \sqrt{\frac{p(1-p)}{n}}$

- Hypothesis Testing (P of 0.05 is usually used)
  - State Hypothesis

    $H_a: \boldsymbol{\mu} \neq \boldsymbol{\mu_o}$ (Two sided Hypothesis)

    $H_a: \boldsymbol{\mu} > \boldsymbol{\mu_o}$- (One sided Hypothesis)

    $H_a: \boldsymbol{\mu} < \boldsymbol{\mu_o}$- (One Sided Hypothesis)

    Use z tables, but t distribution if n<20

$$Z = \frac{\bar{X} - \mu_o}{SE(x)} \qquad Z = \frac{\bar{X} - \mu_o}{SE(p)}$$

$$t = \frac{\bar{X} - \mu_o}{SE(x)}$$

Errors in Hypothesis testing

	H_o	
	True	False
Accept	$\mathbf{1-\alpha}$ **(confidence interval)**	Type II error (β)
Reject	Type 1 error (α)	$\mathbf{1-\beta}$**(power)**

For a given sample size (n), lowering α increases β. Probability of a type II error decreases with increases in n.

3. Comparing two means and proportions

Is there a statistically significant difference? (Sample size>20)

A. Means

1. **Confidence Interval**

$$CI = (\bar{X}_1 - \bar{X}_2) \pm Z_{\frac{a}{2}} \times \sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}$$

2. **Significance testing – State hypothesis first.**

$$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

Interpret the Z score

Is there a statistically significant difference? (Sample size<20) → Use t distribution (Additional Assumption is that there is commonality of variance)

Then

1. **Confidence Interval**

$$CI = (\bar{X}_1 - \bar{X}_2) \pm t^* \times \sqrt{S_P^2(\frac{1}{n_1} + \frac{1}{n_2})}$$

Where

$$S_P^2 = \frac{(n_1 - 1)\, S_1^2 (n_2 - 1) S_2^2}{(n_1 + n_2 - 2)}$$

2. **Significance test**
   a. **Unpaired t- test (Assumes commonality of variance)**

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{S_P^2(\frac{1}{n_1} + \frac{1}{n_2})}}$$

   b. **Welch t-test (If the variances are different_ Unequal Variances)**

(i) **First: Test for equality of variances using F-test**

F-test

$$F_{(v1,v2)} = \frac{S_1^2}{S_2^2}$$

$$df\ for(V1 = n1 - 1), (V2 = n2 - 1)$$

If the test is significant do the welch t-test

(ii) **Second: Welch t-test** (Resembles the Z test)

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2})}}$$

   c. **What if the data is paired?** Observations are not independent, however different pairs are independent

(i) **Paired t-test**

Steps

1. Calculate the differences between the observations on each pair (Distinguish positives and negatives)
2. Calculate the mean difference $\bar{d}$
3. State the hypothesis
4. Calculate standard deviation of the differences S_d
5. From the above, calculate the **Standard Error of the Difference** $SE(\bar{d}) = \frac{S_d}{\sqrt{n}}$
6. Calculate t statistic

$$t = \frac{\bar{d}}{SE(\bar{d})}$$

7. Interpret t value. Compare your value to the critical t_{n-1} distribution → this will give the p-value for the paired t test.

B. Proportions

1. Confidence Interval

$$\mathrm{CI} = \ \mathrm{p\ difference} \pm 1.96\ \times \mathrm{SE\ (difference)}$$

$$\mathrm{SE(difference)} = \ \sqrt{\left[\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right]}$$

2. Statistical Significance

$$Z = \frac{(p_1 - p_2)}{\sqrt{\bar{p}\ \bar{q}\ (\frac{1}{n_1} + \frac{1}{n_2})}}$$

Where $\bar{p} \ = \frac{X_1+X_2}{n_1+n_2}$

4. Association of two categorical variables

1. **Chi-square test**

   Assumptions – Sample size>40 & smallest expected value ≥ 5

   If n is between 20& 40 & smallest expected value is at least 5

   (i) Cross tabulate – Exposures and Outcomes

	Disease +	Disease -	
Factor +	a	b	(a+b)
Factor -	c	d	(c+d)
	(a+c)	(b+d)	n

   (ii) State the hypothesis

   (iii) Chi-square formulae

$$x^2 = \sum \frac{(\mathrm{O} - \mathrm{E})^2}{\mathrm{E}}$$

$$E = \frac{\text{row total x column total}}{\text{Overall total}}$$

   (iv) Establish degree of freedom = (r-1)x(c-1)

   (v) Check the Chi-square matching that degree of freedom at 95% confidence interval $x^2_{0.05}$

   (vi) Compare the two: if less than critical value → accept the null hypothesis

2. **If the above assumptions don't apply – use** Fisher's exact test.(Only for 2 by 2 tables)

   Steps

   (i) Cross Tabulate exposures and outcomes

Exposure	Disease +	Disease -	(**Outcomes**)
Factor +	a	b	(a+b)
Factor -	c	d	(c+d)
	(a+c)	(b+d)	n

   (ii) State the hypothesis

(iii) Do Fisher's exact test

$$p = \frac{\binom{a+c}{a} + \binom{b+d}{d}}{\binom{n}{a+b}}$$