Principles of Inference: a prelude to hypothesis

Goal: - Primary objectives of a statistical analysis is to use data from a sample to make inferences about the population from which the sample was drawn.

The sampling distributions discussed earlier in this lecture note (χ^2, t, F) play a pivotal role in this aspect.

Statistical Inference

- a statement (hypothesis) about the value of the parameter (that we are interested in estimating about a population by using a random sample)
- a measure of the reliability of that statement in terms of a probability (that the decision is incorrect)

Hypothesis Testing

A hypothesis usually results from speculation concerning observed behavior, natural phenomena or established fact. If the hypothesis is stated in terms of population parameters such as the mean & variance, the hypothesis is called a Statistical hypothesis. Data from a sample are used to test the validity of the hypothesis.

Principles of Hypothesis Testing

- Null & Alternate hypothesis. (H_0) (H_1)
- Decision Making / Rejection (Critical) region
- Errors in Decision
  - Types of error α, β. (Type I, Type II)

5-Step procedure for Hypothesis Testing.

Step ① :- Specify H_0, H_1 & an acceptable level of α.

Step ② :- Define a sample based test statistic (eg. $\bar{X}$, s^2 etc) & the rejection region for H_0.

Step ③ :- Collect the sample data & calculate the test statistic.

(Combo step)

Step ④ :- Make a decision to either Reject or fail to reject H_0.

Step ⑤ :- Interpret the result in the language of the problem (i.e. provide Confidence Interval (C.I.), type & probability of error for the kind of decision being made).

Application of $Z \sim N(0,1)$ D"

Example (Quality control in a packaging)

A company that packages salted peanuts in 8kg jars is interested in maintaining control on the amount of peanuts put in jars by one of its machines. Control is defined as averaging 8 kg per jar & not consistently over/underfilling the jars. To monitor this control, a sample of 16 jars is taken from the packaging line at random time intervals & their contents weighed. The mean wt. of peanuts in these 16 jars (test statistic) will be used to test the null hypothesis that the machine is indeed working properly. If it is found not to be doing so, a costly adjustment will be needed.

Step ①

Hypothesis $\begin{cases} H_0: \mu = 8 \\ H_1: \mu \neq 8 \end{cases}$; set α

Step ② & ③ Test statistic $\longrightarrow \bar{X} = \sum_{i=1}^{n} X_i / n$

Step ④ Rejection Region (Critical) $\longrightarrow$ eg. $\{\bar{X} < 7.9 \text{ or } \bar{X} > 8.1\}$

Step ⑤ Type of error

Type (I) :- When H_0 is actually true & our analysis based on our test statistic rejects H_0.

Type (II) :- H_0 is actually not true & our inference based on Test Statistic fails to reject H_0

	In The Population	
The Decision	H_0 is True	H_0 is NOT True
H_0 is NOT rejected	Decision is Correct	type II error (w/ probability β)
H_0 is rejected	Type I error (w/ probability α)	Decision is correct

Calculating α

$$\alpha = P(\bar{X} < 7.9 \text{ or } \bar{X} > 8.1 \text{ when } \mu = 8)$$

Assume (for now) that we somehow know σ of the population (a guess) to be 0.2.

$n = 16$ (This will often not be the case & the n may be req'd for analysis)

$$P(\bar{X} < 7.9) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{7.9 - 8}{0.2/\sqrt{16}}\right)$$

$$= P(Z < -2.0) = 0.0228$$

and

$$P(\bar{X} > 8.1) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{8.1 - 8}{0.2/\sqrt{16}}\right)$$

$$= P(Z > 2.0) = 0.0228$$

$$\therefore \alpha = P(\bar{X} < 7.9) + P(\bar{X} > 8.1)$$

$$= 0.0228 + 0.0228 = 0.0456.$$

$$= \text{Prob}(\text{Type I error})$$

∴ Probability of adjusting the m/c when it doesn't need it is slightly less than 0.05 (5%)

Calculating β

This is not always straightforward.
A type I error is committed when H_0 is rejected if $\bar{X} \neq 8$ even though $\mu = 8$

The sampling D^n for the test statistic ($\bar{X}$) is different for different values of $\mu \neq 8$.

For practical purpose, several different values for β is obtained by considering a handful of μ values eg. {7.80, 7.90, 7.95, 8.05, 8.10, 8.02}

Each of this μ value will give a range of $\bar{X}$ for a globally fixed α.

These β vs μ gives us an "Operating characteristic (OC)" curve

Power of a test = $(1-\beta)$; power vs μ is called power curve.

Reading assignment:

Importance of power of a test.

** There is always a tussle between α & β.
reducing β will inc. α & vise-versa
but generally α is more important to control (reduce) at the expense of a reasonable β.

The significance level of a test is the maximum acceptable probability of rejecting a true NULL hypothesis (i.e. max: allowable type I error probability).

→ This will enable us to define a rejection region

disadv of significance level.

i) One may not have a clear idea of what an appropriate max^m allowable type I error probability should be for a given test statistic

ii) significance level may also be sensitive to minor changes in sample statistic.

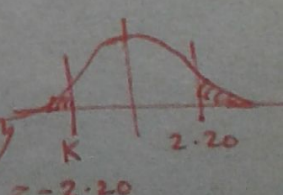
Alternatively, a method of reporting the results of a significance test w/o having to choose an exact level of significance, but instead leave that decision to the individual who will actually act on the conclusion of the test — p-value report.

eg. If $\bar{X} = 7.89$ is obtained

$$P\left(z < \frac{7.89-8}{0.4/\sqrt{16}}\right) = P(z < k) = P(z < -2.2)$$

$P(|z| > 2.20) \xrightarrow[\text{of } N(0,1)]{\text{Symmetry}} 2P(z > 2.20)$

$= 0.0278$



∴ We can say that the probability of observing a test statistic at least this extreme if H_0 is true is 0.0278.

Truth (Actual Situation)

Decision	H_0 True	H_0 false
Do NOT reject H_0	Correct decision w/ prob $1-\alpha$	Incorrect Decision β ← Type II error
Reject H_0	In Correct Decision w/ prob α ← Type I error	Correct Decision $1-\beta$

α = Prob (Type I error)

β = Prob (Type II error)

Application of t-Distribution in hypothesis testing

Two sample test for mean

$X_1 \sim N(\mu_1, \sigma^2)$

$X_2 \sim N(\mu_2, \sigma^2)$

} Note σ^2 is same for both (samples) R.V.

$X_{1i} \sim N(\mu_1, \sigma^2)$; $i = 1, 2, \ldots, n_1$

$X_{2i} \sim N(\mu_2, \sigma^2)$; $i = 1, 2, \ldots, n_2$

Step ①:

H_0 : $\mu_1 = \mu_2$

H_1 : $\mu_1 \neq \mu_2$ or $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$

Also, select α level of significance.

Step ② & ③: Test Statistic

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $S^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1 + n_2 - 2)}$; S_i^2 is sample variance of i^{th} group.

Under H_0:

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

Step ④ Rejection (Critical region)

Reject H_0 in favor of $H_1 (\mu_1 \neq \mu_2)$ if

$$|t| \geq t(\alpha/2, n_1 + n_2 - 2)$$

Reject H_0 in favor of $H_1 (\mu_1 > \mu_2)$ if

$$t \geq t(\alpha, n_1 + n_2 - 2)$$

Reject H_0 in favor of $H_1 (\mu_1 < \mu_2)$ if

$$t \leq -t(\alpha, n_1 + n_2 - 2)$$

example

In comparison of two gasoline brands, a consumer survey reveals the following:

→ A full tank of Brand (A) requires 4 cans and covers 546 kms w/ a std. deviation of 31 kms.

→ A full tank of Brand (B) requires 4 cans & covers 492 kms w/ a std. deviation of 26 kms.

Assume that both populations (A) & (B) are sampled from Normal D^{n}s w/ equal variances. Test

$H_0 : \mu_1 = \mu_2$

vs

$H_1 : \mu_1 > \mu_2$ at 0.05 level of significance

Soln :-

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 > \mu_2$

$\alpha = 0.05$

$\bar{X}_1 = 546, \; S_1 = 31, \; n_1 = 4$

$\bar{X}_2 = 492, \; S_2 = 26, \; n_2 = 4$

$$S^2 = \frac{(4-1)31^2 + (4-1)26^2}{4+4-2}$$

$\Rightarrow S = 28.609$

$$\therefore t(\text{under } H_0) = \frac{546 - 492}{28.609\sqrt{\frac{1}{4}+\frac{1}{4}}} = 2.67$$

$\sim t(0.05, 4+4-2)$

from the table of t-D^n :-

$t(0.05, 6) = 1.943$ (one-tail t-D^n)

$= t(0.05, 6)$

$\therefore t = 2.67 > t(0.05, 6) \Rightarrow$ Reject H_0 in favour of H_1. #

Analysis of Variance

The t-statistic test for 2 means cannot be generalized for multiple-groups/population means. For such a case, we require Anova.

Data :-

y_{ij}, $i = 1, \ldots, t$; $j = 1, \ldots, n_i$ } i.e. t groups, Σn_i observations

If $n_i = n \; \forall \; i \Rightarrow$ "Balanced" data.

Null hypothesis :- $H_0 : \mu_1 = \mu_2 = \cdots = \mu_t$

H_1 : atleast 1 equality is not satisfied

Assumption :- Each group is distributed $N(\mu_i, \sigma^2)$

Note σ^2 is same across population groups.

Organization of Data

Factor Levels	Observations	Totals	Means	Sum of sqs.
1	$y_{11}\ y_{12} \cdots y_{1n_1}$	$Y_{1\cdot}$	$\bar{y}_{1\cdot}$	SS_1
2	$y_{21}\ y_{22} \cdots y_{2n_2}$	$Y_{2\cdot}$	$\bar{y}_{2\cdot}$	SS_2
⋮	⋮	⋮	⋮	⋮
i	$y_{i1}\ y_{i2} \cdots y_{in_i}$	$Y_{i\cdot}$	$\bar{y}_{i\cdot}$	SS_i
⋮	⋮	⋮	⋮	⋮
t	$y_{t1}\ y_{t2} \cdots y_{tn_t}$	$Y_{t\cdot}$	$\bar{y}_{t\cdot}$	SS_t
Overall		$Y_{\cdot\cdot}$	$\bar{y}_{\cdot\cdot}$	SS_P

Factor level totals

$$Y_{i.} = \sum_j y_{ij}$$

Factor level means

$$\bar{y}_{i.} = \frac{Y_{i.}}{n_i}$$

Overall total $Y_{..} = \sum_{i,j} y_{ij}$

Overall mean $\bar{y}_{..} = \frac{Y_{..}}{\sum_i n_i}$

The computation of estimated variance forms through the calculations below

$$SS_i = \sum_j (y_{ij} - \bar{y}_{i.})^2 \quad ; \quad i = 1, 2, \ldots, t$$

$$\text{OR} \quad \sum_j y_{ij}^2 - \frac{(Y_{i.})^2}{n_i}$$

Pooled sum of sq

$$SS_p = \sum_i SS_i$$

Pooled degrees of freedom $\sum n_i - t$

$$s_p^2 = \frac{SS_p}{\sum_i n_i - t} = \frac{\sum_i SS_i}{\sum_i n_i - t}$$

If the individual variances are available s_i^2,

then $s_p^2 = \frac{\sum_i (n_i - 1) s_i^2}{\sum_i n_i - t}$

$$s^2_{\text{means}} = \frac{\sum_i (\bar{y}_{i.} - \bar{y}_{..})^2}{(t-1)}$$ is the variance estimate of the factor level means.

Under the null hypothesis & By the 1st th^m. of sampling D^n; the factor level means have D^n w/ mean μ & variance $= \sigma^2/n$

sided trial variance of samples at each factor level is the same, σ^2.

$$s^2_{means} = \sigma^2/n \Rightarrow n s^2_{means} = \sigma^2 \text{ w/ } (t-1) \text{ d.o.f}$$

An alternate estimate of σ^2 is s_p^2 w/ $t(n-1)$ d.o.f (here $n_i = n \forall i$)

Recall from the def^n. of F; D^n; the $F-D^n$ represents the ratio of 2 independent estimates of a common variance.

$$\therefore F_{calculated} = \frac{n s^2_{means}}{s_p^2} > F_\alpha(t-1, t(n-1))$$

then reject H_0.

example:- An experiment to compare the yield of 4 varieties of rice was conducted. Each of 16 plots on a test-farm where soil fertility was fairly homogeneous was treated alike relative to H_2O & fertilizer. Four plots were randomly assigned each of the 4 varieties of rice. The yield in kg/acre was recorded for each plot for this randomized experiment. Do the data presented in the following table indicate a difference in the mean yield bet'n the 4 varieties?

Table of rice yields

variety	Yields				$Y_{i\cdot}$	$\bar{y}_{i\cdot}$	SS_i
1	934	1041	1028	935	3938	984.50	10085.00
2	880	963	924	946	3713	928.25	3868.75
3	987	951	976	840	3754	938.50	13617.00
4	992	1143	1140	1191	4466	1116.50	22305.00
Overall					15871	991.94	49875.75

Soln :-

Hypothesis $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

H_1 : not all varieties have the same mean

Where μ_i is the mean yield per acre for variety i.

$$n s^2_{mean} = n \sum_i (\bar{y}_{i.} - \bar{y}_{..})^2 / (t-1) \quad ; \quad \begin{matrix} n=4 \\ t=4 \end{matrix}$$

$$= 4\{(984.5 - 991.94)^2 + \cdots\cdots + (1116.50 - 991.94)^2\} / (4-1)$$

$$= 29977.06$$

$$s_p^2 = \sum_i SS_i / t(n-1) = (10,085 + \cdots + 22,305) / 4 \times 3 = 49875.75 / 12$$

$$= 4156.31$$

$$F_{cal} = \frac{29977.06}{4156.31} = 7.21$$

$\alpha = 0.01$ (say set apriori by statistician)

$$F_{0.01}(3, 12) = 5.95$$

d.o.f in numerator $= (t-1) = 3$

d.o.f in denominator $t(n-1) = 12$

$\therefore F_{cal} > F_{0.01}(3, 12)$

Reject H_0!

i.e Difference may exist in the yields of 4 varieties.

Computational convenience for calculating F-Statistic

Between group sum of squares,

$$SSB = \sum_i \frac{Y_{i.}^2}{n_i} - \frac{Y_{..}^2}{\sum n_i} \quad \text{w/ } dfB = (t-1)$$

W/in group sum of squares,

$$SSW = \sum_{i,j} y_{ij}^2 - \sum_i \frac{Y_{i.}^2}{n_i} \quad \text{w/ } dfW = \sum_i n_i - t$$

Total sum of squares,

$$TSS = SSB + SSW \quad \text{w/ } dfT = dfB + dfW = \sum_i n_i - 1$$

1-way Analysis of Variance table

Source	df	SS	$MS = \frac{SS}{df}$	F
Bet'n groups	$t-1$	SSB	MSB	$\frac{MSB}{MSW}$
W/in groups	$\sum_i n_i - t$	SSW	MSW	
Total	$\sum_i n_i - 1$	TSS		

ANOVA table for our earlier example

Source	df	SS	MS	F
Bet'n varieties	3	89,931.19	29977.06	7.21
W/in varieties	12	49,875.75	4156.31	
Total	15	139,806.94		

Same concl as before