RNR / ENTO 613 _ --Inferences Using t-Distributions

Introducing t-tools with permutation tests

The fundamental goal of statistics: to draw inferences on (unknown) population parameters using sampling data.

Because the data obtained from sampling are inherently variable, we need to find a way to separate variability due to the sampling process from real patterns in the data.

One way to do this is to set a **null hypothesis** that provides a basis to estimate the variability introduced by the sampling process. The actual statistic (s) computed from the sampling data can then be compared to the variability due to sampling expected if the null hypothesis is correct.

Permutation (or randomization):

Resampling techniques (e.g. permutation) provide a direct way to estimate the expected variability associated with the sampling process.

Ex: Is there a difference in weight between females and males in the class?

To answer that question, we obtain a sample of 5 males and 5 females and calculate the difference between the average weight of those samples:

Wf = 120 lbs; Wm = 160 lbs; Wm – Wf = 40 lbs. ** It would seem that males are heavier than females.

We want to reach a reasonably probable conclusion based on a single sampling event (an "experiment"), but we do not know whether the observed difference (40 lbs) is meaningful, given that the sampling process always introduce variation that is ultimately contained in the statistic computed from samples.

One way to solve this problem is to set a null hypothesis. A typical hypothesis would be:

Ho: Wm – Wf = 0

If the null hypothesis is correct, i.e. *if males and females* ***are from the same population****, or in other words if males and females* ***do have similar weights***, being a female or male is irrelevant with respect to weight. We can thus resample the pooled data, each time allocating the 10 observations at random to the category "female" and "male", with 5 observations falling in each category.

Thus, resampling the data under the Ho of no difference, we obtain something like this……

Resampling event	"Wf"	"Wm"
1	130	141
2	151	135
3	122	179
4	127	172
Etc	Etc	Etc

Resampling many times and each time calculating the statistic Wm –Wf, *we generate the expected pattern of variation in Wm – Wf under the null hypothesis of no difference between males and females*. This pattern of variation in the statistic **depends** on size of the samples *and* on the true variability in weight (σ) in our population.

We can now compare the statistic computed **from the 2 samples we got** to the distribution generated by the permutation process. If the statistic computed from the 2 samples is unusually large, there are 2 possibilities:

a) we obtained a non-representative set of samples (i.e. Ho *was correct*: by chance, we sampled 5 "extreme" males or females). This would occur with a probability P, if the null hypothesis is correct.
b) males are indeed heavier than females (i.e. Ho *was not correct*)

Rejecting Ho in a) is a type 1 error
Rejecting Ho in b) is appropriate

<<Sleuth Display 1.8 >>

t-tools:

T-tools for comparison of two population means use the same general framework as above:

a) set a null hypothesis (generally Ho: no difference between population means)
b) obtain 2 samples (a single "experiment") and calculate t-ratio
c) look at the expected distribution of the statistic (*t-ratio*) pertaining to the difference between the 2 population means if the null hypothesis is true (i.e.,f statistical tables)
d) compare the t-ratio calculated from sampling data with its expected distribution under Ho
e) reject or do not reject Ho

The differences between t-tools and permutation:

a) the statistic used for inferences is a **t-ratio** rather than the difference between sample averages.

b) the distribution of the t-ratio expected, for the sampling process used and if the null hypothesis is true, is provided by statistical tables (Student's t-distribution).

The reason we use a t-ratio is historical: we **can** predict theoretically the distribution of a t-ratio [the output of those predictions are Statistical Tables], but we can only predict the distribution of the difference between 2 sample averages empirically [e.g. with permutations]). They had no computers at the beginning of the last century, so only the first option was practical….

<<Sleuth Display 2.10>>

Type of inferences with t-tests

1. **Hypothesis testing:** Use sample data to determine if a population mean, the difference between paired observations, or the difference between 2 population means, differ from a hypothesized value. We test the null hypothesis that μ or $\mu_1 - \mu_2$ is equal to a hypothesized value (μ_0).

   Type of Statistical finding: *The left hippocampus volume was smaller in twins with schizophrenia than in twins without schizophrenia (paired t-test, two-sided p-value = 0.0061)*

2. **Confidence intervals:** Use sample data to estimate plausible values for a population mean, the difference between paired observations, or the difference between 2 population means.

   Type of Statistical finding: *The mean hippocampus volume is estimated to be between 0.07 and 0.33 cm^3 larger (95% confidence interval) in twins without schizophrenia than in twins with schizophrenia.*

Practical Context

One-sample t-test: Draw inferences on a single population mean. Example: A producer wants to estimate the mean weight of turkeys in her stock because turkeys of a given size sell well.

Paired t-test: Test whether a population mean difference between *paired observations* differs from a previously decided value (often zero). Example: Hippocampus size in 15 pairs of schizophrenic or healthy twins (Sleuth).

Two-sample t-test: Draw inferences on the difference between two population means. Example: Humerus length of house sparrows that died or survived in a storm (Sleuth).

One-sample Hypothesis

Context: Test the null hypothesis that the population mean (μ) is equal to some specified value (μ_0).

Example: What is the average size of turkeys in a flock?

A producer has a flock of 3500 turkeys. She wants to know whether those turkeys weigh about 25 lbs, which is the size preferred by consumers.

Ho: $\mu o = 25$
Ha: $\mu o \neq 25$

One sample-hypothesis testing using the Z-Ratio

If the population standard deviation (σ) was known (which is rarely the case), hypothesis tests about population means could be based on the **standard (Z) normal distribution.**

For any normally distributed population of observations, a standard normal distribution is obtained by applying a Z transformation:

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

The sampling distribution of such Z-ratios is called **standard normal** because it reduces **any** normal distribution (specified by μ and σ) to a **unique** standard normal distribution, which is centered on 0 and has a standard deviation of 1.

<<FIG 1>>

To test whether $\mu = \mu_o$, we ask how likely we are to observe the measured deviation between $\bar{Y}$ (our best estimate of μ) and μ_o. The answer depends on the value of the Z-ratio obtained from sampling data.

The Z-ratio of a sample average for a 1-sample hypothesis is:

$Z = \dfrac{\bar{Y} - \mu_o}{\sigma/\sqrt{n}} = \dfrac{\bar{Y} - \mu_o}{SD(\bar{Y})}$. (we use SD here because we know the population variance: later with t-tools we will use SE).

Where $\sigma/\sqrt{n}$ is the standard deviation of sample averages.

<<Fig 2.4, Sleuth>>

If a Z-ratio was calculated for many random samples from a single population, we would obtain a standard normal distribution of the Z-ratio (**$\mu = 0$, $\sigma = 1$) only when the null hypothesis is correct** (that is, when the sampled population is correctly described by the null hypothesis).

The greater the difference between μ_o and μ, the more extreme will be the Z-ratios calculated from sample data.

<< FIG 2 >>

When the null hypothesis is correct, we expect a Z-ratio obtained from sampling data to be close to zero, **i.e. within the range delimited by the chosen critical values with a probability 1 - α** (e.g. if we choose $\alpha = 0.05$, 95 % of the Z-ratios should fall within ± 1.96).

If the Z-ratio is more extreme than the chosen critical value, we reject the statement that null hypothesis is correct.

One sample-hypothesis testing using the t-Ratio

In most cases, the value of σ is not known, so the **Student's t-distribution** must be used.

The Student's t-distribution was developed by W. S. Gosset, an employee of the Guinness brewery, who published under the pseudonym of "Student".

The t-ratio is very similar to the z-ratio:

For a 1-sample hypothesis is: $t = \dfrac{\bar{Y} - \mu_o}{\mathrm{SE}(\bar{Y})}$, where $\mathrm{SE}(\bar{Y}) = \dfrac{s}{\sqrt{n}}$

The difference between the t- and z-ratio, and thus between the Z- and the Student's t-distribution, is that there is **an error associated with estimating SE ($\bar{Y}$) when σ is not know**.

The smaller the sample size, the **lower** is the precision of SE ($\bar{Y}$). The t-distribution was derived to compensate for this uncertainty in estimation of SE ($\bar{Y}$).

The shape of the t-distribution approaches that of the Z-distribution as n increases. If n $\rightarrow\infty$, there is no error in estimation of SE ($\bar{Y}$), so the t- and Z-distributions become identical.

<< FIG 3>>

So the shape of the Student's t-distribution depends on the **degree of freedom (i.e. sample size)** used to estimate the unknown standard error of sample averages.

****A rule to determine degrees of freedom associated with standard errors is to subtract from the sample size, n, the **number of parameters being estimated** (*df = n – no. parameter estimated*). In tests about one mean, SE is the only parameter estimated (only one mean in the model), so d.f. = n –1.

For the turkey example, the producer decided to weigh 20 individuals **randomly obtained** from the stock. From this sample, we calculate:

Sample average, $\bar{y}$ = 22.85: $\frac{\sum_{i=1}^{n} y_i}{n}$

Sample variance, s^2 = 4.45: $\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$

Standard error, SE ($\bar{Y}$) = SQRT (4.45/20) = 0.47: $\sqrt{\frac{s^2}{n}}$.

We establish a rejection criterion (α level) and associated critical value:

1- The traditional way is to determine a critical value for the test, given the *df* and chosen α-level (say α= 0.05). From a **t-table**, we find the **critical t-value.** For *df* = 20-1 = 19 and α= 0.05, the critical values are ± 2.093, for a 2-tailed test.

In this example, $t = \frac{22.85 - 25.0}{0.47} = -4.558$, with 19 *df*.

Compare the critical value of the test statistic to the observed value of the test statistic: $t_{obs} < t_{crit}$, (- 4.558 < - 2.093), so **reject the null hypothesis for the alternative** at α = 0.05.

2- With computers it is now easy to determine the exact probability associated with the t-ratio calculated under the assumption that the null hypothesis is correct. For t_{19} = - 4.558, the exact P = 0.0002.

(Interpretation: we expect a t-ratio this or more extreme only 2 times in 10,000 *if the null hypothesis is correct*).

Results: the null hypothesis that the turkeys weigh on average 25 lbs is unlikely to be true. They can still grow for a while….

Hypothesis tests are useful to **guess** whether a population parameter is different from a specified value, but they do not allow precise estimation of population parameters. In this example, the test does not tell us how much the actual average weight differs from the targeted weight. **Confidence intervals** can do this.

A confidence interval for the mean

Confidence intervals provide answers to the general question:

What are plausible values for the (population) parameter of interest?
In this example, *What are the plausible values for the mean weight of the turkeys?*

The general formula to transform a sample mean into a t-ratio is:

$$\text{t-ratio} = \frac{\bar{Y} - \mu}{SE(\bar{Y})}$$

We know that most of the sample averages calculated from sampling data **will fall near the true value of the population mean (i.e. μ)**.

Similarly, **most t-ratio obtained from sampling data will fall near 0**. Specifically, such t-ratios would conform to a Student's t-distribution and fall (1 - α) 100 % of the time within some chosen critical values denoted $\pm t_{df}(1-\alpha/2)$.

<< FIG Sleuth p. 35>>

In our example, the critical values which would include 95% of the most likely t-ratios with 19 d.f. and α= 0.05 are –2.093 and 2.093.

Thus 95 % of the time that a sample is drawn from our study population, we expect the resulting "incompletely specified" t-ratios to fall within those critical values:

$$-2.093 \leq \frac{22.85 - \mu}{0.47} \leq 2.093$$

We can thus solve this equation to find a confidence interval for μ

$$22.85 - 2.093 \times 0.47 \leq \mu \leq 22.85 + 2.093 \times 0.47$$

$$21.86 \leq \mu \leq 23.84$$

So the formula for a 100 (1 - α) % CI for the mean is:

$\bar{Y} - t_{df}(1-\alpha/2)\ SE(\bar{Y}) \leq \mu \leq \bar{Y} + t_{df}(1-\alpha/2)\ SE(\bar{Y})$

or

$\bar{Y} \pm t_{df}(1-\alpha/2)\ SE(\bar{Y})$

<<OUTPUT FROM JMP>>

A confidence interval defines the interval that encompasses μ with a 100 (1 - α) % confidence. The lower bound for μ is called the *lower confidence limit*, while the upper bound is called the *upper confidence limit*.

For a 95 % confidence interval, the critical value [$t_{df}(1-\alpha/2)$] is close to 2 when sample size gets over 20 (for an infinite n, it is 1.96, as in the z-distribution). As a rough guess, we therefore expect a confidence interval to extend ± 2 SE on each side of $\bar{Y}$.

The appropriate way to interpret confidence intervals is:

If all possible samples of size n were drawn from a population and a 95 % confidence interval was calculated for each sample, 95% of those intervals would contain the true population mean (μ).

Hypothesis test and **confidence intervals** are therefore closely related. For a specified level of confidence (α), a confidence interval specifies plausible values for μ. If those values do not overlap with the value specified by the null hypothesis, it is safe to reject the null hypothesis.

Paired-Sample Hypotheses

> *Context:* Data from one sample are related to data from another sample (i.e. we have **paired observations** not independent from each other). Each pair comprises the same individual tested twice, or two individuals expected to be similar for genetic or environmental reasons. The statistical model for analysis must take into account such non-independence.
>
> *Approach:* Determine if the average of the difference between paired observations differs from some value (often 0). Formally, test the null hypothesis that the population mean difference between pairs is equal to 0 (or some other value) with a paired t-test.

Example: Is the concentration of a plant phytochemical different between the upper and lower part of the root?

Measure chemical concentration in the upper and lower part of the root in 10 plants.

Reduce paired observations to a **single set of differences** for subsequent analysis. The model reduces to a 1-sample t-test, comparing the sample of differences to the hypothesized value.

LOWER	UPPER	DIF (= UP – LOW)
10	12	2
11	13	2
10	11	1
7	10	3
12	12	0
7	9	2
11	14	3
8	12	4
8	10	2
8	8	0

Define the mean difference between pairs, μ_d : The population parameter to estimate. State the two-tailed null hypothesis as:

Ho: $\mu_d = 0$
Ha: $\mu_d \neq 0$

Estimate μ_d with $\bar{d}$. Calculate the t-statistic as: t = $\dfrac{\bar{d} - 0}{SE(\bar{d})}$

The only difference between a 1-sample t-test and a paired t-test is that n in the paired case is the number of **differences** rather than the total number of observations.

With n =10, df = 10 – 1 = 9

$\bar{d}$ =1.9 ng, s^2 = 1.65 ng^2, s = 1.29 ng, SE($\bar{d}$) = 0.407

t_{obs} = 1.9/0.407 = 4.67.

Critical value at $\alpha = 0.05$: $t_{9,\ 0.975} = 2.26$. Since t_{obs} (4.67) > t_{crit} (2.26), reject the null hypothesis for the alternative.

The exact two-tailed P-value for this t-statistic is P = 0.0012.

The formula for a 100 (1- α) % confidence interval for the mean difference is:

$\bar{d} \pm t_{df}(1-\alpha/2)\ SE(\bar{d})$

For $\alpha = 0.05$ (i.e. a 95 % CI):

$1.9 \pm (2.26)(0.407) = 1.9 \pm 0.92 = (0.98, 2.82)$

This CI excludes the value stated by the null hypothesis (i.e. 0). Thus a null hypothesis test would be rejected at the stated $\alpha = 0.05$.

Results: The concentration of the plant phytochemical differed between the upper and lower part of the root (two-sided p-value =0.0012, paired t-test). The 95 % CI for the estimated mean difference ($\bar{d}$ =1.9 ng, SE = 0.407) between the concentration in the top and lower part is from 0.98 to 2.82.

Two-sample Hypotheses

Context: Data collected from two independent samples, perhaps using a completely randomized design for 2 groups.

Approach: Determine if population means differ from each other (or from some other value). Formally, test the null hypothesis that the population means are the same using a 2-sample t-test.

Example: Compare the heights or two groups of plants that were treated with water only (n_1 =10) of with a fertilizer (n_2 = 8).

State null and alternate hypotheses as:

Ho: $\mu_1 = \mu_2$
Ha: $\mu_1 \neq \mu_2$

Or equivalently

Ho: $\mu_1 - \mu_2 = 0$
Ha: $\mu_1 - \mu_2 \neq 0$

The value specified in the null hypothesis is usually 0 but can be specified to be any ecologically meaningful difference.

Test statistic for comparing means from two independent samples:

$$t = \frac{(\bar{Y}_1 - \bar{Y}_2) - \mu_o}{SE(\bar{Y}_1 - \bar{Y}_2)}$$

SE ($\bar{Y}_1 - \bar{Y}_2$) is the pooled standard error for the difference between averages from the two independent samples.

That standard error is:

$$\text{SE}\,(\overline{Y}_1 - \overline{Y}_2) = \sqrt{\left(\frac{\sigma_1^2}{n1} + \frac{\sigma_2^2}{n2}\right)}$$

If we assume the 2 populations have the same variance (an assumption of the test), we obtain

$$\text{SE}\,(\overline{Y}_1 - \overline{Y}_2) = \sigma\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

To estimate the population standard deviation, we use a *weighted average* of the sample variances (i.e. s_p below). A weighted average is used because precision of the sample variance depends on sample size: a weighted average is more precise than a simple average to estimate σ.

So s_p, the **pooled estimate of standard deviation** is**:**

$$s_p = \sqrt{\frac{(n1-1)s_1^2 + (n2-1)s_2^2}{(n1+n2-2)}} \qquad \text{d.f} = n1 + n2 - 2$$

Finally, we substitute s_p for σ, which yields:

$$\text{SE}\,(\overline{Y}_1 - \overline{Y}_2) = s_p \sqrt{\left(\frac{1}{n1} + \frac{1}{n2}\right)} \qquad \text{d.f} = n1 + n2 - 2$$

With this estimate of SE, we can calculate the t statistics for comparing means from 2 independent samples.

> *Example*: Compare heights of 2 groups of plants which received water (n1 = 10) or a fertilizer (n2 = 8).

State the 2-tailed null hypothesis:

> Ho: $\mu_1 = \mu_2$
> Ha: $\mu_1 \neq \mu_2$

Calculate $\overline{Y}_1 = 51.91$ $s_1^2 = 11.36$ n =10

$\overline{Y}_2 = 56.55$ $s_2^2 = 9.88$ n = 8

$s_p = \sqrt{\frac{(9X11.36)+(7X9.88)}{10+8-2}} = 10.71 \text{ cm}^2$

$SE(\overline{Y}_1 - \overline{Y}_2) = 10.71\sqrt{\left(\frac{1}{10}+\frac{1}{8}\right)} = 1.55 \text{ cm}$

The appropriate t-statistic:

$$t = \frac{(\overline{Y}_1 - \overline{Y}_2) - \mu_o}{SE(\overline{Y}_1 - \overline{Y}_2)} = \frac{(51.91 - 56.55) - 0}{1.55} = \frac{-4.64}{1.55} = -2.99$$

If we set $\alpha = 0.05$, the critical value for this test, with 10 + 8 – 2 =16 df is - 2.12

Because t_{obs} (- 2.99) < t_{crit} (- 2.12), reject the null hypothesis for the alternative.

The exact P-value for $t_{16} = 2.99$ is P = 0.0087

A **two-tailed** test is used when the difference between means could go either way. Sometimes we are only interested in a directional test, i.e in a **one-tailed** hypothesis. For example, there are situations where we know that the population parameter can only be greater (or smaller) than the value specified by the null hypothesis.

For a one-tailed test, the test statistics is the same; the alternative hypothesis is stated differently, and the p-value is exactly ½ of that for a 2 tailed-test.

For the fertilizer case, if we have **good a priori reasons to believe** that the plants that received fertilizer can only grow taller than the ones that received water:

Ho: $\mu_1 = \mu_2$
Ha: $\mu_1 < \mu_2$

Or equivalently:

Ho: $\mu_1 - \mu_2 = 0$
Ha: $\mu_1 - \mu_2 < 0$

For $\alpha = 0.05$, $t_{crit} = -1.746$. This value is **smaller** than for an equivalent 2-tailed test, because now we consider as extreme the 5 % of t-ratios that are smaller than t_{crit}. For the 2-tailed test, we considered the 2.5 % of the t-ratios that fell either to the left or right of the critical value.

Because t_{obs} = - 2.99 is smaller than the one-tailed critical value (- 1.746), reject the null hypothesis as being correct.

The exact 1-tailed p-value for t_{16} = - 2.99 is P = 0.0043 (exactly ½ of the p-value for the 2-tailed test).

To get a better idea of the difference between population means, we derive a confidence interval:

$$\overline{Y}_1 - \overline{Y}_2 \pm t_{df}\,(1-\alpha/2)\ SE\ (\overline{Y}_1 - \overline{Y}_2)$$

For the plant height example,

df = 16
$\overline{Y}_1 - \overline{Y}_2$ = 51.91-56.55=- 4.64
SE ($\overline{Y}_1 - \overline{Y}_2$) = 1.55 cm
$t_{df}\,(1-\alpha/2)$ = 2.12

The 95% CI for the difference μ1 - μ2 is:

51.9 –56.55 ± (2.12)(1.550) = - 4.64 ± 3.286 = -7.926, -1.354.

Interpretation for such 95% CI is:

If all possible pairs of random samples were taken from both populations of size n1 and n2, and a new CI was generated each time, then 95 % of those CI would capture the true value of μ1 - μ2.

Note that zero is not included in the 95% CI for the difference between group means. This implies that the null hypothesis μ1 - μ2 = 0 would have been rejected at α = 0.5.

Results: Addition of the fertilizer **caused** an increase in plant growth compared to plants receiving water (t_{16} = 2.99, P = 0.0087). The 95% CI for the estimated mean difference in growth for plants receiving water ($\bar{y}$ = 51.91 cm) or the fertilizer ($\bar{y}$ = 56.55 cm) was 1.35 to 7.93.

***Note: the 95% CI for a 1-tailed test is calculated as follow:

$$\overline{Y}_1 - \overline{Y}_2 \pm t_{df}\,(1-\alpha)\ SE\ (\overline{Y}_1 - \overline{Y}_2)$$