Two-Sample Inferences

In Chapters 6 & 7, you learnt how to test a particular value of a parameter (proportion p or mean μ) in one population.
Sometimes, you don't compare one population to a particular value of its parameter (p_0 or μ_0), but two populations together

Ex:
- Average SAT score of men VS women in UCSD
- Average height of San Diegans VS New-Yorkers
- Average age of husbands & wives
- Average income of children compared to their parents

In all these examples, you have 2 samples coming from 2 populations

$X_1, \dots, X_n$: Population 1

$Y_1, \dots, Y_m$: Population 2

⚠ n and m have no reason to be equal in general.

9.0 Paired Samples

Sometimes though, the members of the 2 populations have some direct link where each member of one population is paired with a member of the other. In such a case, $m = n$.

Exc: • Weight Pre-diet / Weight Post-diet

• Husband Age / Wife Age

We have $X_1, \dots, X_n \overset{i.i.d}{\sim} N(\mu_X, \sigma_X^2)$

$Y_1, \dots, Y_n \sim N(\mu_Y, \sigma_Y^2)$ (all parameters unknown)

If we want to test $H_0: \mu_X - \mu_Y = 0$, (i.e $\mu_X = \mu_Y$)

we just consider the differences

$$\boxed{\Delta_i = X_i - Y_i} \quad (1 \leq i \leq n)$$

and we perform a 1 sample t-test on $\Delta_1, \dots, \Delta_n$

Exc: You decide to research global warming in the U.S. You choose 62 random cities and look up the high temperature on Jan 1st 1970 and Jan 1st 2018.

Clearly, the data are paired: hot locations will have high readings in both 1970 and 2018. Similarly for cold locations.

You calculate $\Delta = Temp_{2018} - Temp_{1970}$ for each location and find the differences Δ_i; have $\bar{\Delta}_{62} = 1.1°F$ and

$$S_\Delta^2 = \frac{1}{62-1} \sum_{i=1}^{62} (\Delta_i - \bar{\Delta}_{62})^2 = 24.1$$

If we care about the true (unknown) average temperature rise $\mu_{2018} - \mu_{1970}$ in the U.S., our confidence interval would be

$$CI_{1-\alpha} = \left[\bar{\Delta}_{62} - t_{1-\frac{\alpha}{2}, 62-1} S_\Delta / \sqrt{64} \; , \; \bar{\Delta}_{64} + t_{1-\frac{\alpha}{2}, 62-1} \frac{S_\Delta}{\sqrt{n}} \right]$$

$$\overset{\alpha = 5\%}{=} [-0.144°F, 2.346°F].$$

Rk: Nothing new statistically speaking, just the "difference trick".

In what follows, we'll only focus on <u>unpaired data</u>, meaning that the X_i's and Y_j's are independent.

9.2 Testing $\mu_X = \mu_Y$

We focus here on tests with $H_0 : \mu_X - \mu_Y = 0$. The alternative would be $>$, $<$ or $\neq$ depending on the situation but the method is the same.

The variances σ_X^2 and σ_Y^2 are not known, but the method we'll use will differ when $\sigma_X^2 = \sigma_Y^2$ and when $\sigma_X^2 \neq \sigma_Y^2$. We'll learn later (Chapter 9.3) how to actually test $H_0 : \sigma_X^2 = \sigma_Y^2$.

<u>Equal Variances</u>

Thm: Let $\begin{cases} X_1, \dots, X_n \overset{iid}{\sim} N(\mu_X, \sigma_X^2) \\ Y_1, \dots, Y_m \overset{iid}{\sim} N(\mu_Y, \sigma_Y^2) \end{cases}$ be independent with $\sigma_X^2 = \sigma_Y^2 = \sigma^2$

Then $$T = \frac{(\overline{X}_n - \overline{Y}_m) - (\mu_X - \mu_Y)}{\sqrt{\frac{1}{n} + \frac{1}{m}}\, S_p} \sim t_{n+m-2}$$ (t-distribution)

where $S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 + \sum_{j=1}^{m}(Y_j - \overline{Y}_m)^2}{n+m-2}$.

pooled

(5)

Proof: One easily checks that $Q = \dfrac{\bar{X}_n - \bar{Y}_m - (\mu_x - \mu_y)}{\sqrt{\frac{1}{n} + \frac{1}{m}}}$ has distribution $N(0, \sigma^2)$.

Furthermore, if $W^2 = (n-1) S_x^2 + (m-1) S_y^2$, we notice that:

$$\begin{cases} \cdot \dfrac{(n-1)S_x^2}{\sigma^2} \sim \chi^2_{n-1} \\ \cdot \dfrac{(m-1)S_y^2}{\sigma^2} \sim \chi^2_{m-1} \\ \cdot S_x^2 \perp\!\!\!\perp S_y^2 \quad \text{since the } X_i\text{'s and } Y_j\text{'s are independent.} \end{cases}$$

(See Review on Gaussian vectors)

Hence, $\dfrac{W^2}{\sigma^2} \sim \chi^2_{n+m-2}$.

Finally, $W \perp\!\!\!\perp Q$ (Same), so that $T = \sqrt{n+m-2}\,\dfrac{Q}{W} \sim t_{n+m-2}$.

Rk:

- The assumption that $\sigma_x = \sigma_y = \sigma$ is key in the proof and cannot be released. It allows to compute a pooled variance S_p^2, and hence, improves the power of the tests associated to this result.
- If $n+m-2 \geq 100$, you may use the quantiles of the η-table, since they get super close to those of the t-tables with $df \geq 100$.

Unequal variances

Thm (Welch) let $\begin{cases} X_1, \dots, X_n \overset{iid}{\sim} N(\mu_X, \sigma_X^2) \\ Y_1, \dots, Y_m \overset{iid}{\sim} N(\mu_Y, \sigma_Y^2) \end{cases}$ be independent

Then

$$T = \frac{(\bar{X}_n - \bar{Y}_m) - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \overset{\text{Approximately has distribution}}{\approx} t_\nu ,$$

where $$\nu = \frac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2}{\frac{S_X^4}{n^2(n-1)} + \frac{S_Y^4}{m^2(m-1)}}$$

Proof: (rough)

$$T = \frac{\left(\frac{\bar{X}_n - \bar{Y}_m - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}\right)}{\sqrt{\frac{V}{\nu}}} \overset{?}{\approx} \frac{N(0,1)}{\chi_{df}}$$

$$\sqrt{\frac{V}{\nu}} = \left(\frac{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}\right)$$

Need to find the best ν = df for this approximation.

(7)

To find the best approximation (in distribution) of

$$\frac{\frac{S_x^2}{n} + \frac{S_y^2}{m}}{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}} \quad \text{by} \quad \frac{V}{\nu} \quad \text{where } V \sim \chi^2_\nu,$$

we equate the first two moments of both sides, and get

$$\nu = \frac{\left(\frac{S_x^2}{n} + \frac{S_y^2}{m}\right)^2}{\frac{S_x^4}{n^2(n-1)} + \frac{S_y^4}{m^2(m-1)}}.$$

Rk:
- This is not a rigourous proof, only a heuristic. What we've shown is that the best value for ν is the announced one, not that it actually does approximate well the distribution.
- In practice, it does provide a good test statistic.
- See "Satterthwaite approximation" for a more general statement of such an approximation by χ^2 distributions.

9.3 Testing $\sigma_X = \sigma_Y$ (the F-test)

As highlighted by the previous Section 9.2, equality of variances helps designing more powerful tests and inference procedures. Also, variances provide an informative parameter that may describe populations. In what follows, we describe a test statistic to decide whether $\sigma_X = \sigma_Y$ or not, from data.

First, let's describe a family of distributions that'll help in the process.

Def: (Fisher distribution)

Given $d_1, d_2 \in \mathbb{N}$, we denote by F_{d_1, d_2} the distribution of the random variable

$$\frac{U_1/d_1}{U_2/d_2}, \text{ where } \begin{cases} \cdot\ U_1 \sim \chi^2_{d_1} \\ \cdot\ U_2 \sim \chi^2_{d_2} \\ \cdot\ U_1 \perp\!\!\!\perp U_2 \end{cases}$$

F_{d_1, d_2} is called the Fisher distribution with (d_1, d_2) d.f.

An immediate consequence of Cochran yields the following

(9)

Thm: let $\begin{cases} X_1, \dots, X_m \overset{iid}{\sim} N(\mu_X, \sigma_X^2) \\ Y_1, \dots, Y_m \overset{iid}{\sim} N(\mu_Y, \sigma_Y^2) \end{cases}$ be independent.

Then

$$\frac{\sigma_X^2}{\sigma_Y^2} \times \frac{S_Y^2}{S_X^2} \sim F_{m-1, m-1}$$

Proof: From Cochran, $\begin{cases} \cdot (m-1) S_X^2 \sim \chi^2_{m-1} \\ \cdot (m-1) S_Y^2 \sim \chi^2_{m-1} \end{cases}$, and since the X_i's and Y_j's are independent, so are $(m-1)\frac{S_X^2}{\sigma_X^2}$ and $(m-1)\frac{S_Y^2}{\sigma_Y^2}$, which proves the result. ∎

Rk:
- This allows to test $H_0: \sigma_X = \sigma_Y$ without even knowing (nor having inferred) μ_X and μ_Y.
- Depending on the alternative hypothesis, the rejection region has 1 or 2 pieces.

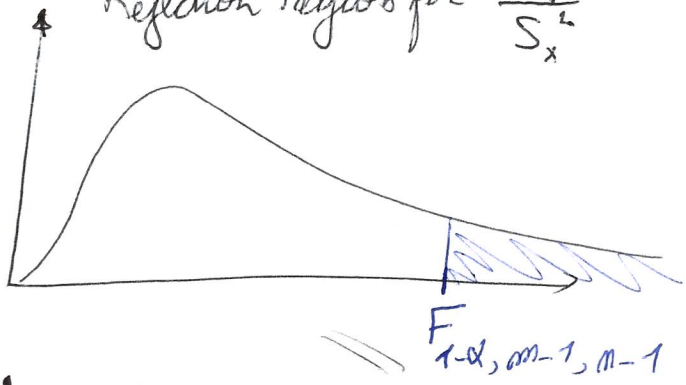
In all the cases the test statistic to use is

$$F = \frac{S_Y^2}{S_X^2}$$

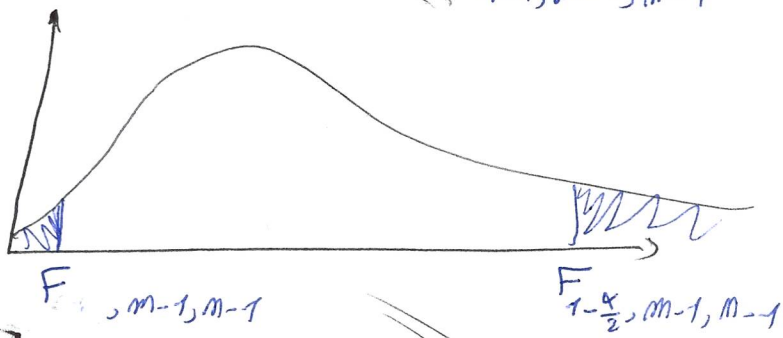
(or equivalently $\frac{S_X^2}{S_Y^2}$)

Rejection region for $\frac{S_Y^2}{S_X^2}$

$H_1: \sigma_X < \sigma_Y$



$H_1: \sigma_X \neq \sigma_Y$



$H_1: \sigma_X > \sigma_Y$

- For testing means, the F-test is often used to determine if variances are equal

1) State hypotheses $H_0: \mu_X = \mu_Y$, $H_1: \mu_X \left\{ \begin{matrix} < \\ \neq \\ > \end{matrix} \right\} \mu_Y$

2) Check the assumptions
   - Paired data?
   - Normality?
   - Independence?
   - Equal or not-equal variances? ⟶ Perform a F-test

3) Perform the right t-test.

9.4 Binomial Data: Testing $H_0: p_X = p_Y$

Back to categorical data, we now want to infer equality of the proportions p_X and p_Y of a trait in two populations.

Data then take the form of two sequences of 0-1 variables

$$\begin{cases} X_1, \dots, X_n \overset{iid}{\sim} \text{Bernoulli}(p_X) = \mathcal{B}(p_X) \\ Y_1, \dots, Y_m \overset{iid}{\sim} \text{Bernoulli}(p_Y) = \mathcal{B}(p_Y) \end{cases}$$

Thm: If n and m are large enough so that $\begin{cases} np_X,\ n(1-p_X) \geq 10 \\ mp_Y,\ m(1-p_Y) \geq 10 \end{cases}$

then

$$\frac{(\hat{p}_X - \hat{p}_Y) - (p_X - p_Y)}{\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{m}}} \approx N(0,1)$$

Where $\hat{p}_X = \frac{X_1 + \dots + X_n}{n} = \overline{X}_n$ and $\hat{p}_Y = \frac{Y_1 + \dots + Y_m}{m} = \overline{Y}_m$

Proof:

From the C.L.T. $\begin{cases} \sqrt{n}(\hat{p}_X - p_X) \xrightarrow[n\to\infty]{\mathcal{D}} N(0,\ p_X(1-p_X)) \\ \sqrt{m}(\hat{p}_Y - p_Y) \xrightarrow[m\to\infty]{\mathcal{D}} N(0,\ p_Y(1-p_Y)) \end{cases}$

(12) Hence, since $\hat{p}_x$ and $\hat{p}_y$ are independent,

$$(\hat{p}_x - \hat{p}_y) - (p_x - p_y) \overset{\sim}{\sim} N\left(0, \frac{p_x(1-p_x)}{n} + \frac{p_y(1-p_y)}{m}\right)$$

Hence,

$$\frac{(\hat{p}_x - \hat{p}_y) - (p_x - p_y)}{\sqrt{\frac{p_x(1-p_x)}{n} + \frac{p_y(1-p_y)}{m}}} \xrightarrow[\Rightarrow]{n,m\to\infty} N(0,1).$$

But from the law of large number, $\begin{cases} \hat{p}_x \xrightarrow[n\to\infty]{a.s.} p_x \\ \hat{p}_y \xrightarrow[m\to\infty]{a.s.} p_y \end{cases}$, so that

from Slutsky's theorem,

$$\frac{(\hat{p}_x - \hat{p}_y) - (p_x - p_y)}{\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n} + \frac{\hat{p}_x(1-\hat{p}_y)}{m}}} \xrightarrow[n,m\to\infty]{\Rightarrow} N(0,1)$$

Rk: - The conditions $np_x, mp_y, n(1-p_x), m(1-p_y) \geq 10$ are here to ensure the convergence result holds. Indeed, if one of these conditions don't hold, the associated Bernoulli is skewed and the rate in the C.L.T. get really slow (See Berry-Esseen Thm).

(13) . This result provides a confidence interval for $p_x - p_y$:

$$\left[(\hat{p}_x - \hat{p}_y) \pm z^* \sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n} + \frac{\hat{p}_y(1-\hat{p}_y)}{m}} \right], \text{ where } z^* = N(0,1)_{1-\frac{\alpha}{2}}$$

For testing $p_x = p_y$, one should use another statistic than the one above. Indeed, under the null, $p_x = p_y = p$, so that

$$\sqrt{\frac{p_x(1-p_x)}{n} + \frac{p_y(1-p_y)}{m}} = \sqrt{\left(\frac{1}{n} + \frac{1}{m}\right) p(1-p)} .$$

As a consequence, we can use another estimator of $p_x = p_y = p$ to estimate the variance term. This would result in a better approximation in Slutsky, and hence a more powerfull test.

We let $\hat{p}_e = \dfrac{X_1 + \cdots + X_n + Y_1 + \cdots + Y_m}{n+m} = \dfrac{n\hat{p}_x + m\hat{p}_y}{n+m}$

(14)

Prop: If $p_x = p_y = p$, and that $(n+m)p \geq 10$, $(n+m)(1-p) \geq 10$, then

$$\frac{\hat{p}_x - \hat{p}_y}{\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)\hat{p}_e(1-\hat{p}_e)}} \approx N(0,1)$$

Proof: Same as before, except you use that $\hat{p}_e \longrightarrow p$ in Slutsky.

Rk: • You see that the condition $(n+m)p, (n+m)(1-p) \geq 10$ is weaker than the one previously. Actually, that'll provide a better power to the test that use this statistic

• When to pool / not to pool ?

* Confidence Interval : Do not pool
* Test : Pool

the reason for this is that for a test, you do you reasoning under $H_0 : p_x = p_y = p$ but not for C.I's.

- When performing a test for proportions, keep in mind to check that you actually are working with a Binomial model:
  - counts derived from n Bernoulli trials
  - Number of trials fixed in advance
  - Probability of success constant across trials

9.5 Confidence intervals for two-sample problems

As for one-sample problems, the knowledge of the distributions of the above statistics (and others) provides confidence intervals.

Make sure you can write C.I's of level α in all the cases below:

Parameter of interest	Case	
$\mu_X - \mu_Y$	• σ_X and σ_Y known	
	• σ_X and σ_Y unknown	• $\sigma_X = \sigma_Y$
		• $\sigma_X \neq \sigma_Y$
$\frac{\sigma_X^2}{\sigma_Y^2}$	• μ_X and μ_Y known	
	• either μ_X or μ_Y known	
	• neither μ_X nor μ_Y known	

Can you do it for $\frac{\sigma_X}{\sigma_Y}$ too?