# On choosing appropriate hypothesis tests

Vong Jun Yi (vongjy.github.io)

## Variances

- Population variance: The variance of the entire population, denoted by $\sigma^2$
- (Biased) Sample variance: The variance of a particular sample from a population, denoted by $S^2$
- Unbiased estimator for population variance: An estimate of population variance when it is unknown, denoted by $s^2$ or $\hat{\sigma^2}$.
- **Bessel's correction** - The variances obey the following relationship:

$$s^2 = \frac{n}{n-1} S^2$$

- Standard error refers to $\frac{s}{\sqrt{n}}$.

## Means

- Population mean: denoted by $\mu$.
- Sample mean: usually denoted by $\overline{x}$.
- $\overline{x}$ is an unbiased estimator for $\mu$.

## Parameters and tests for confidence intervals

- Suppose a significance level of $\alpha \in (0,1)$, then $p := 1 - \frac{\alpha}{2}$,

| Case | Test | Confidence interval |
|---|---|---|
| Population mean with known population variance | $z$-test | $\overline{x} \pm z_p \frac{\sigma}{\sqrt{n}}$ |
| Population mean using large sample (unknown $\sigma^2$) | $z$-test | $\overline{x} \pm z_p \frac{s}{\sqrt{n}}$ |
| Population mean using small sample (unknown $\sigma^2$) | $t$-test | $\overline{x} \pm t_{p,n-1} \frac{s}{\sqrt{n}}$ |
| Population proportion, $\hat{p}$ (large sample) | $z$-test | $\hat{p} \pm z_p \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ |
| Difference in population means using small sample | $t$-test | $(\overline{x} - \overline{y}) \pm t_{p,n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ |
| Difference in population means using large sample | $z$-test | $(\overline{x} - \overline{y}) \pm t_{p,n_1+n_2-2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ |

| Case | Test | Confidence interval |
| --- | --- | --- |
| Difference in population means with matched pairs | $t$-test | $\bar{d} \pm t_{p,n-1} \dfrac{s_d}{\sqrt{n}}$ |

## Hypothesis testing (Difference in means)

| Tests | Assumptions |
| --- | --- |
| Two-sample $t$-test | - Underlying distributions are normal.<br>- Populations are independent.<br>- Population variance of the two populations is the same (but may be unknown). |
| Two-sample $z$-test (Normal distribution) | - Underlying distributions are normal.<br>- Large sample sizes.<br>- Populations are independent.<br>- Population variance of the two populations is the same (but may be unknown). |
| Paired sample $t$-test | - Differences are normally distributed.<br>- Population variance of the two populations is the same (but may be unknown).<br>- Data are matched pairs (repeated measures design). |

### Two sample $t$-test

- If $n_1$ and $n_2$ are small ($< 30$), and the two populations are normally distributed with an **unknown common variance**, then the test statistic $t$ has the distribution

$$(\overline{X} - \overline{Y}) \sim t_{n_1+n_2-2} \left( \mu_x - \mu_y, s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \right)$$

and

$$t = \frac{(\overline{x} - \overline{y}) - (\mu_x - \mu_y)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

- If the sample sizes are too small to allow us to use $s_x^2$ and $s_y^2$ are estimators, we need to pool these variances (combine them).
- The pooled estimate of the population variance is

$$\begin{aligned} s_p^2 &= \frac{\sum (x - \overline{x})^2 + \sum (y - \overline{y})^2}{n_x + n_y + 2} \\ &= \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y + 2}. \end{aligned}$$

### Two sample $z$-test (Normal distribution)

- If $n_1$ and $n_2$ are large $(\geq 30)$, then the distribution of $(\overline{X}-\overline{Y})$ is given by

$$(\overline{X}-\overline{Y}) \sim N\left(\mu_x-\mu_y, \frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)$$

  and test statistic is

$$z=\frac{\overline{x}-\overline{y}-(\mu_x-\mu_y)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}.$$

- If the population variance is known, the test statistic is

$$Z=\frac{(\overline{x}-\overline{y})-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}} \sim N(0,1).$$

## Paired sample $t$-test

- The test statistic $t$ has the distribution $D \sim N\left(\mu_d, \frac{s_d^2}{n}\right)$ and $t=\frac{\overline{d}-k}{s_d/\sqrt{n}}$.