STAT 515 fa 2023 Lec 16

Two-sample testing

Karl Gregory

Two-sample testing

Suppose it is of interest to compare the means or proportions of two separate populations, for example

- the mean GPAs of female and male honors college students.
- the proportions of rural versus city dwellers who support a candidate.

Consider drawing samples from both populations, where, if means are concerned, μ_1, σ_1^2 and μ_2, σ_2^2 are the relevant population parameters and $\bar{X}_1, S_1^2$ and $\bar{X}_2, S_2^2$ are their sample estimators, and, if proportions are concerned, p_1 and p_2 are the population parameters and $\hat{p}_1$ and $\hat{p}_2$ are their sample estimators.

Concerning the population means μ_1 and μ_2, we are typically interested in hypotheses of the form

$$\begin{array}{lll} H_0\colon\ \mu_1 \geq \mu_2 & \text{or} \quad H_0\colon\ \mu_1 = \mu_2 & \text{or} \quad H_0\colon\ \mu_1 \leq \mu_2 \\ H_1\colon\ \mu_1 < \mu_2 & H_1\colon\ \mu_1 \neq \mu_2 & H_1\colon\ \mu_1 > \mu_2, \end{array}$$

which we prefer to rewrite as

$$\begin{array}{lll} H_0\colon\ \mu_1 - \mu_2 \geq 0 & \text{or} \quad H_0\colon\ \mu_1 - \mu_2 = 0 & \text{or} \quad H_0\colon\ \mu_1 - \mu_2 \leq 0 \\ H_1\colon\ \mu_1 - \mu_2 < 0 & H_1\colon\ \mu_1 - \mu_2 \neq 0 & H_1\colon\ \mu_1 - \mu_2 > 0, \end{array}$$

respectively, and, concerning the population proportions p_1 and p_2, we are typically interested in hypotheses of the same form:

$$\begin{array}{lll} H_0\colon\ p_1 \geq p_2 & \text{or} \quad H_0\colon\ p_1 = p_2 & \text{or} \quad H_0\colon\ p_1 \leq p_2 \\ H_1\colon\ p_1 < p_2 & H_1\colon\ p_1 \neq p_2 & H_1\colon\ p_1 > p_2, \end{array}$$

which we prefer to rewrite as

$$\begin{array}{lll} H_0\colon\ p_1 - p_2 \geq 0 & \text{or} \quad H_0\colon\ p_1 - p_2 = 0 & \text{or} \quad H_0\colon\ p_1 - p_2 \leq 0 \\ H_1\colon\ p_1 - p_2 < 0 & H_1\colon\ p_1 - p_2 \neq 0 & H_1\colon\ p_1 - p_2 > 0. \end{array}$$

It is convenient to write these hypotheses in terms of the differences $\mu_1 - \mu_2$ and $p_1 - p_2$ because we have at hand the sample differences $\bar{X}_1 - \bar{X}_2$ and $\hat{p}_1 - \hat{p}_2$, with which we will build test statistics.

We may also test hypotheses of the form

$$\begin{array}{lll} H_0\colon \mu_1 - \mu_2 \geq \delta & H_0\colon \mu_1 - \mu_2 = \delta & H_0\colon \mu_1 - \mu_2 \leq \delta \\ H_1\colon \mu_1 - \mu_2 < \delta & H_1\colon \mu_1 - \mu_2 \neq \delta & H_1\colon \mu_1 - \mu_2 > \delta, \end{array} \quad \text{or} \quad \text{or}$$

or

$$\begin{array}{lll} H_0\colon p_1 - p_2 \geq \delta & H_0\colon p_1 - p_2 = \delta & H_0\colon p_1 - p_2 \leq \delta \\ H_1\colon p_1 - p_2 < \delta & H_1\colon p_1 - p_2 \neq \delta & H_1\colon p_1 - p_2 > \delta, \end{array} \quad \text{or} \quad \text{or}$$

for some number δ, which would allow us to conclude, were we to reject H_0, that the difference in parameters is less than, not equal to, or greater than the value δ, for example.

Concerning means

From now on, we will assume that the population variances σ_1^2 and σ_2^2 are unknown, as this is generally the case in practice.

To compare two population means μ_1 and μ_2, we will compare the sample means $\bar{X}_1$ and $\bar{X}_2$. To estimate the difference $\mu_1 - \mu_2$, for example, we will use the quantity $\bar{X}_1 - \bar{X}_2$. The following result about the sampling distribution of $\bar{X}_1 - \bar{X}_2$ will be the starting point for finding a test statistic for hypotheses concerning $\mu_1 - \mu_2$.

Sampling distribution result: Sampling distribution of a difference in sample means

Let $X_{11}, \ldots, X_{1n_1}$ and $X_{21}, \ldots, X_{2n_2}$ be random samples from populations having means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively, and let $\bar{X}_1 = n_1^{-1} \sum_{i=1}^{n_1} X_{1i}$ and $\bar{X}_2 = n_1^{-1} \sum_{i=1}^{n_2} X_{2i}$ be the sample means.

1. If both populations are Normally distributed, then

$$\bar{X}_1 - \bar{X}_2 \sim \text{Normal}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

2. Otherwise we have:

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$ has a distribution more and more like the $\text{Normal}(0, 1)$ dist.

for larger and larger sample sizes n_1 and n_2.

Remark 1. *If $n_1 \geq 30$ and $n_2 \geq 30$, we will assume that the distribution of $\bar{X}_1 - \bar{X}_2$ is well-approximated by the Normal$(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$ distribution.*

Recall that in the single population setting, if $\bar{X}$ has the Normal$(\mu, \sigma^2/n)$ distribution, then $(\bar{X} - \mu)/(s/\sqrt{n})$ has the t-distribution with degrees of freedom $n-1$, and we used this fact to build confidence intervals for μ and to test hypotheses about μ. We will proceed similarly in the two-population setting, starting from Sampling distribution result above.

If conditions are satisfied such that

$$\bar{X}_1 - \bar{X}_2 \sim \text{Normal}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right),$$

then we have

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim \text{Normal}(0, 1).$$

If the values σ_1^2 and σ_2^2 were known to us, we could use this fact to construct confidence intervals for $\mu_1 - \mu_2$ or to test hypotheses about $\mu_1 - \mu_2$. However, since we do not know σ_1^2 or σ_2^2, we must estimate them from the data. This will, as in the single population setting, lead us to using the t-distribution in place of the standard Normal distribution. To proceed, we must consider two cases.

Case 1: σ_1^2 and σ_2^2 believed to be equal

We do not know σ_1^2 or σ_2^2, but it may be reasonable to believe that these two unknown quantities are equal, or at least close to equal, to some common variance σ^2_{common}. If this is the case, then instead of having to estimate σ_1^2 and σ_2^2 separately, we can combine the data from both samples to estimate the common variance σ^2_{common}. An unbiased estimator of σ^2_{common} is given by

$$\begin{aligned} S^2_{\text{pooled}} &= \left[\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)^2 + \sum_{i=1}^{n_2}(X_{2i} - \bar{X}_2)^2\right] / (n_1 + n_2 - 2) \\ &= \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}, \end{aligned}$$

where

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)^2 \quad \text{and} \quad S_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(X_{2i} - \bar{X}_2)^2$$

are the sample variances from the two samples. The following result gives us a way to build a confidence interval for $\mu_1 - \mu_2$ or to test hypotheses about $\mu_1 - \mu_2$ when we estimate $\sigma^2_{\text{common}} = \sigma_1^2 = \sigma_2^2$ using S^2_{pooled}.

Sampling distribution result: Studentized difference in sample means under equal variances

Let $\bar{X}_1$ and $\bar{X}_2$ be the means of random samples from populations with means μ_1 and μ_2, respectively, and common variance σ^2_{common}.

1. If both populations are Normally distributed, then

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{S^2_{\text{pooled}}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$ has the t-distribution with degrees of freedom $n_1 + n_2 - 2$.

2. Otherwise we have:

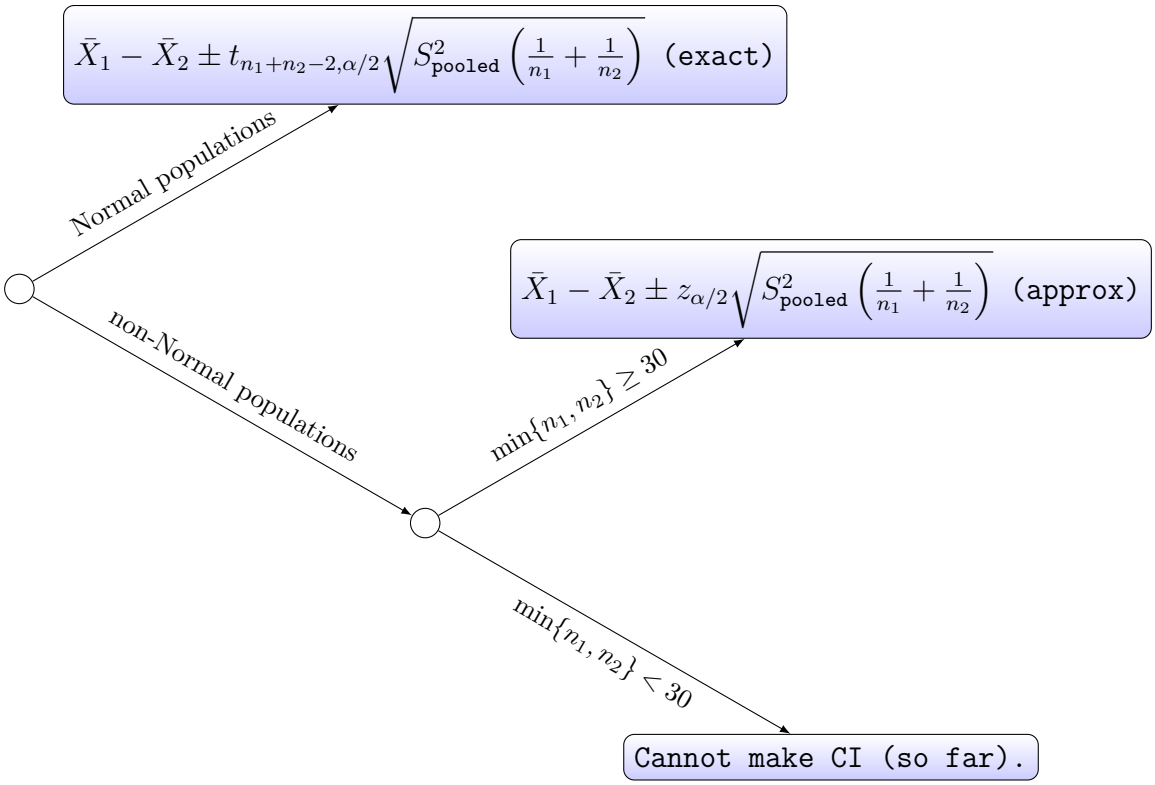
$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{S^2_{\text{pooled}}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$ has a distribution more and more like the $\text{Normal}(0, 1)$ distribution for larger and larger sample sizes n_1 and n_2.

Rule of thumb 1. *If $n_1 \geq 30$ and $n_2 \geq 30$, then even if the populations are non-Normal, we will assume that the sampling distribution of*

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{S^2_{\text{pooled}}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

is well-approximated by the Normal$(0, 1)$ distribution.

We can use the above result to build $(1 - \alpha) \times 100\%$ confidence intervals for $\mu_1 - \mu_2$ in the $\sigma_1^2 = \sigma_2^2$ case according to the following rules:



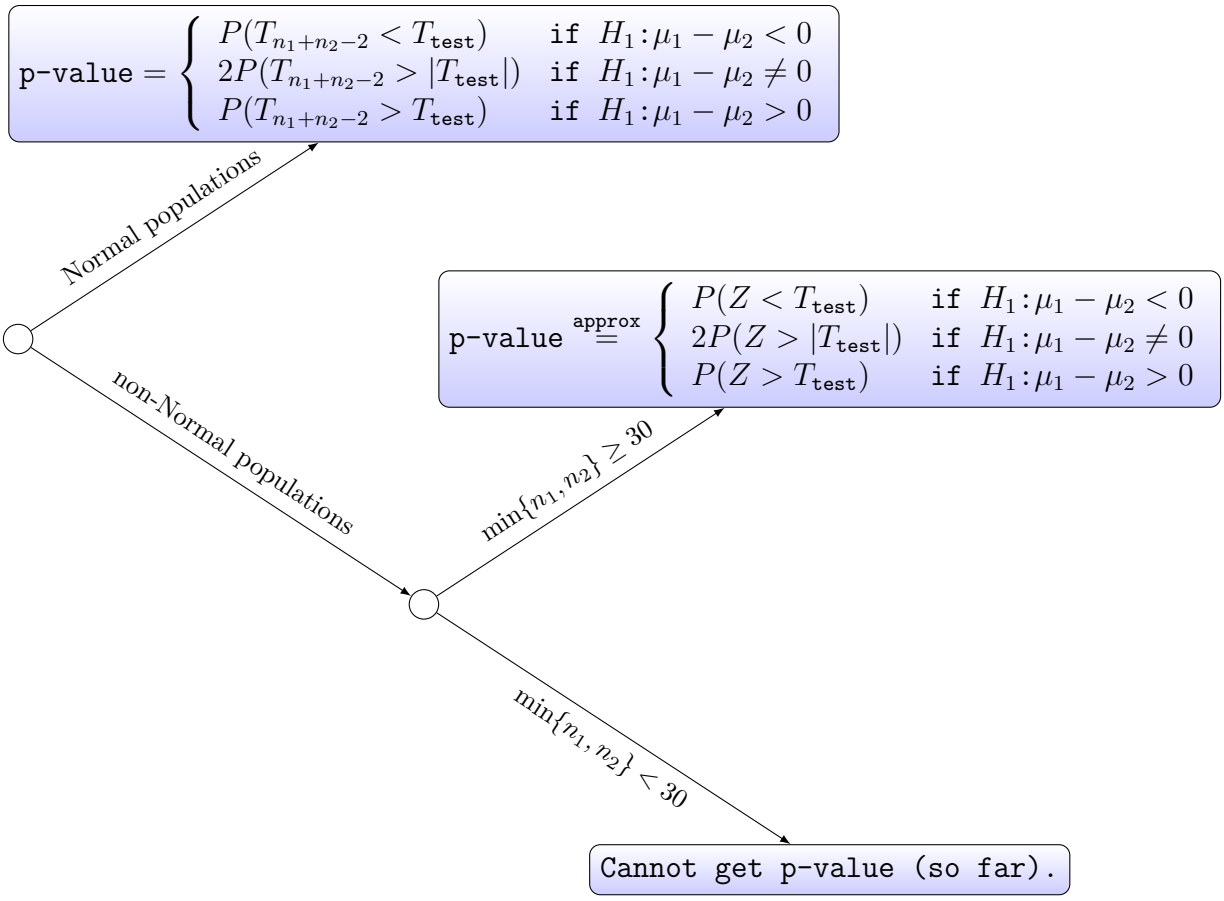
For testing hypotheses of the forms

$$\begin{array}{lll} H_0\text{: } \mu_1 - \mu_2 \geq 0 & \text{or} & H_0\text{: } \mu_1 - \mu_2 = 0 \quad \text{or} \quad H_0\text{: } \mu_1 - \mu_2 \leq 0 \\ H_1\text{: } \mu_1 - \mu_2 < 0 & & H_1\text{: } \mu_1 - \mu_2 \neq 0 \qquad\quad H_1\text{: } \mu_1 - \mu_2 > 0, \end{array}$$

we have the test statistic

$$T_{\text{test}} = \frac{\bar{X}_1 - \bar{X}_2 - (0)}{\sqrt{S^2_{\text{pooled}}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S^2_{\text{pooled}}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}},$$

which is the estimated number of standard deviations $\bar{X}_1 - \bar{X}_2$ lies from $\mu_1 - \mu_2$ when $\mu_1 - \mu_2 = 0$. Note that, as we have done before, we construct the test statistic at the value of the parameter(s) separating the null from the alternative hypothesis. According to Sampling distribution result , we can get a p-value associated with the test statistic T_{test} according to the following rules:



In the diagram, $T_{n_1+n_2-2}$ represents a random variable having the t-distribution with degrees of freedom n_1+n_2-2 and Z represents a Normal$(0,1)$ random variable. The above diagram shows that if the populations are Normal, we will get our p-values using the t-distribution with degrees of freedom n_1-n_2-2; if not, we rely on the Central Limit Theorem, using the z-table to get our p-values so long as n_1 and n_2 are both greater than or equal to 30.

Case 2: σ_1^2 and σ_2^2 believed to be unequal

It may not be reasonable to believe that σ_1^2 and σ_2^2 are equal. In this case we cannot pool the data together to produce a single estimator S^2_{pooled} of a common variance σ^2_{common}. We must estimate σ_1^2 and σ_2^2 by the corresponding sample variances S_1^2 and S_2^2. In this case we have the following result:

Sampling distribution result: Studentized diff. in sample means under unequal variances

Let $\bar{X}_1$ and $\bar{X}_2$ be the means of random samples from populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively.

1. If both populations are Normally distributed, then the sampling distribution of

   $$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$ can be approximated by the t-dist. with degrees of freedom ν^*,

   where

   $$\nu^* = \left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2 \left[\frac{\left(S_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(S_2^2/n_2\right)^2}{n_2 - 1}\right]^{-1}.$$

2. Otherwise we have:

   $$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$ has a distribution more and more like the Normal $(0, 1)$

   distribution for larger and larger sample sizes n_1 and n_2.

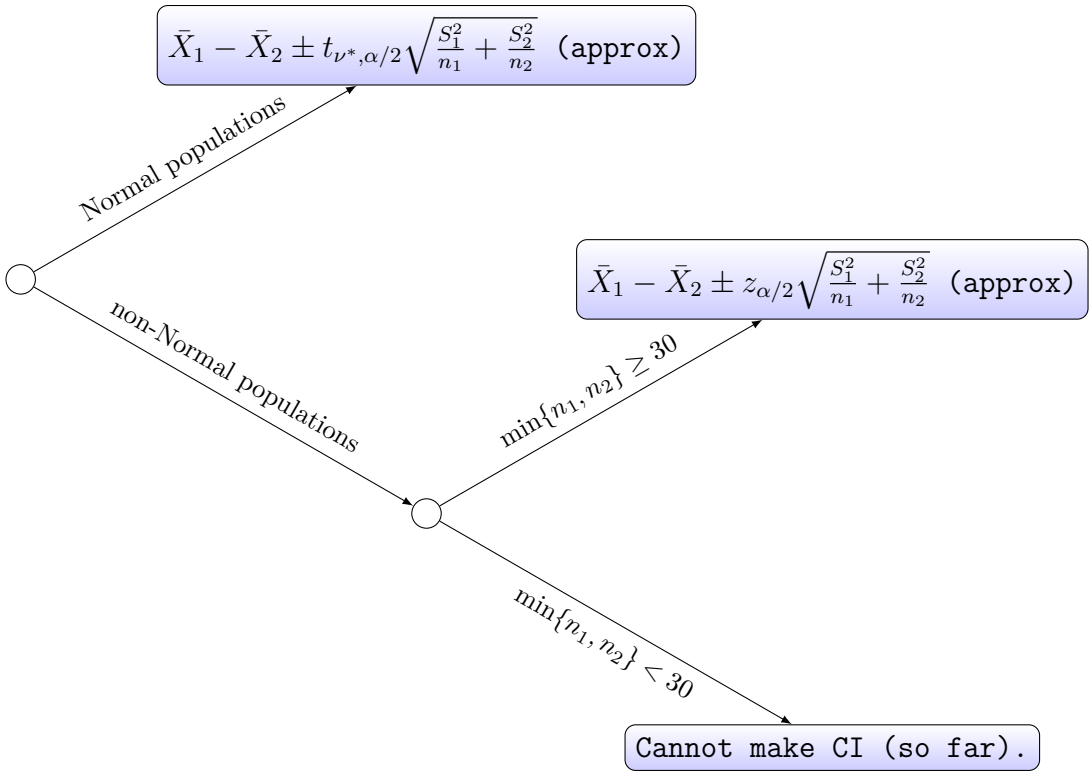
It is worth noting that the value ν^* may not be an integer. Even though we have not discussed t-distributions with a non-integer degrees of freedom, they do exist. If using a table in the book, round ν^* down to the nearest integer below.

Rule of thumb 2. *If $n_1 \geq 30$ and $n_2 \geq 30$, then even if the populations are non-Normal, we will assume that the sampling distribution of*

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

is well-approximated by the Normal $(0, 1)$ *distribution.*

We may use this result to build $(1-\alpha) \times 100\%$ confidence intervals for $\mu_1 - \mu_2$ in the $\sigma_1^2 \neq \sigma_2^2$ case according to the following rules:



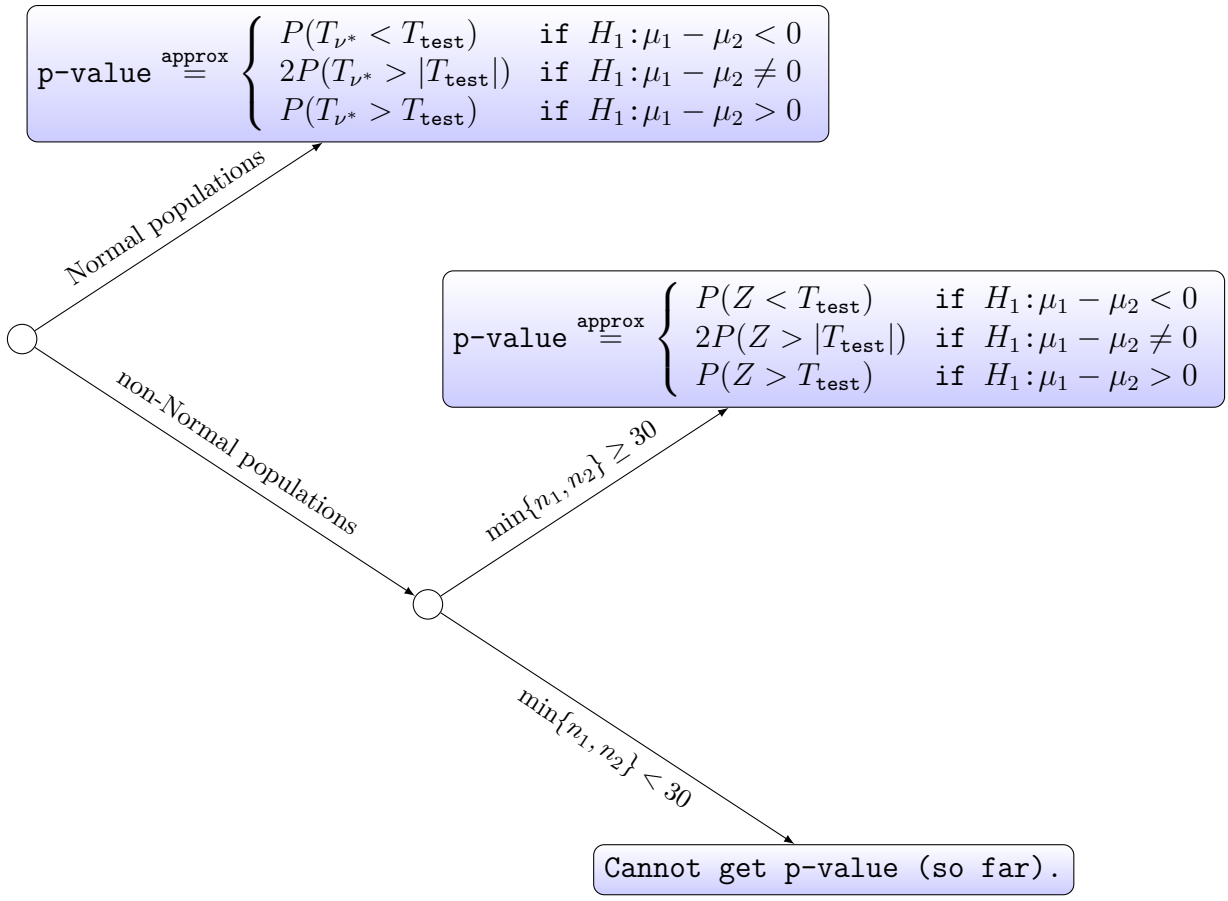
For testing hypotheses of the forms

$$\begin{array}{lll} H_0\colon\ \mu_1 - \mu_2 \geq 0 & \text{or} & H_0\colon\ \mu_1 - \mu_2 = 0 \quad \text{or} \quad H_0\colon\ \mu_1 - \mu_2 \leq 0 \\ H_1\colon\ \mu_1 - \mu_2 < 0 & & H_1\colon\ \mu_1 - \mu_2 \neq 0 \qquad\quad H_1\colon\ \mu_1 - \mu_2 > 0, \end{array}$$

we now have the test statistic

$$T_{\text{test}} = \frac{\bar{X}_1 - \bar{X}_2 - (0)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}},$$

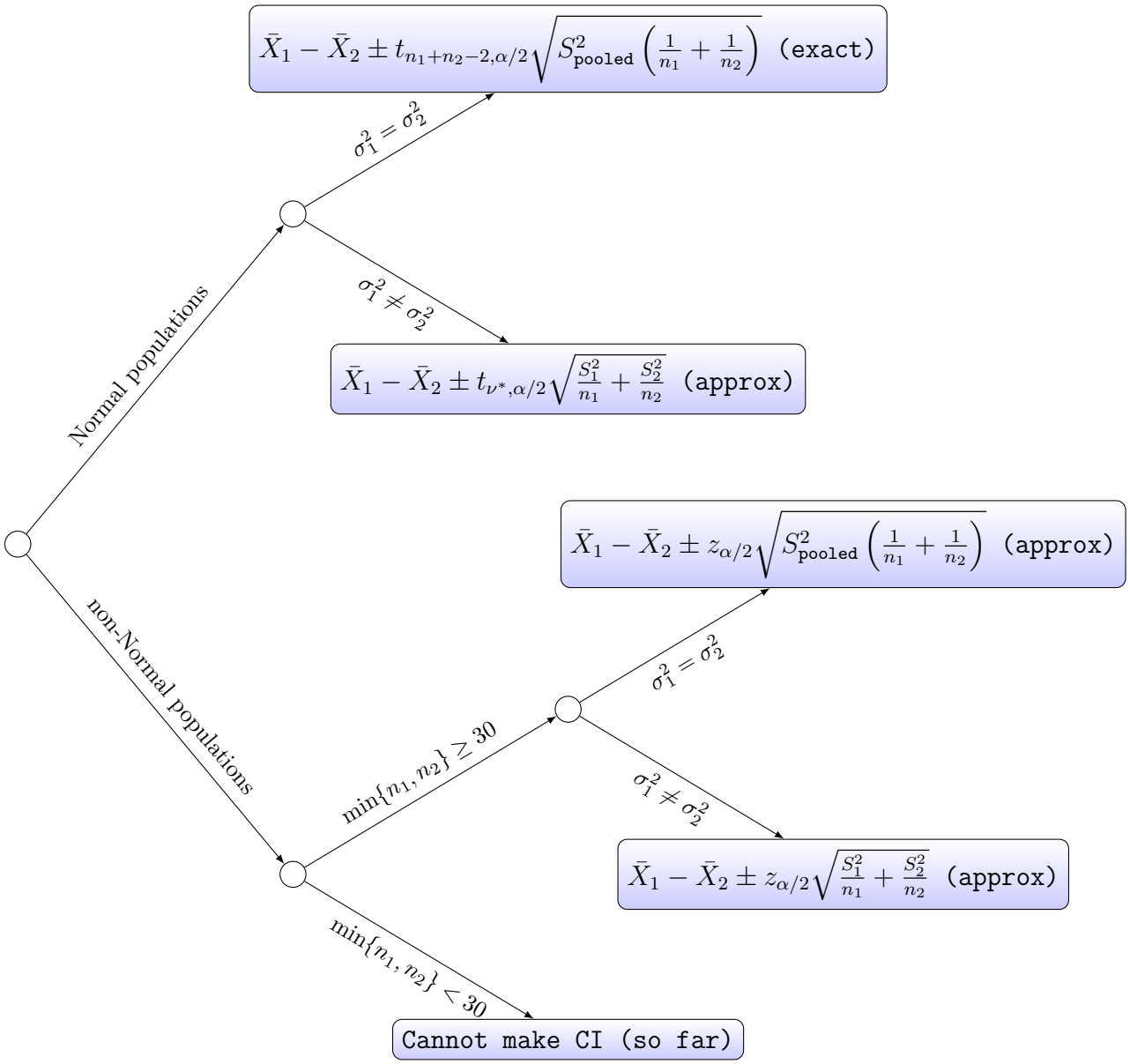
which is the estimated number of standard deviations $\bar{X}_1 - \bar{X}_2$ lies from $\mu_1 - \mu_2$ when $\mu_1 - \mu_2 = 0$. According to sampling distribution result , we can get a p-value associated with T_{test} according to the following rules:



In the diagram, T_{ν^*} represents a random variable having the t-distribution with degrees of freedom ν^* and Z represents a Normal$(0, 1)$ random variable. Similarly to before, the diagram shows that if the populations are Normal, we will get our p-values using the t-distribution with degrees of freedom ν^*; if not, we rely on the Central Limit Theorem, using the z-table to get our p-values so long as n_1 and n_2 are both greater than or equal to 30.

Summary of confidence intervals for differences in means

The diagram below shows how to construct a $(1 - \alpha) \times 100\%$ confidence interval for $\mu_1 - \mu_2$ in different situations:



Exercise. Refer to 9.23 of the textbook. Drug tablets are produced at two sites, and it is of interest whether the two sites produce tablets with the same concentration of the drug. That is, if μ_1 and μ_2 are the mean concentrations of the drug in the tablets produced by sites 1 and 2, respectively, we are interested in testing

$$H_0\text{: } \mu_1 - \mu_2 = 0 \text{ versus } H_1\text{: } \mu_1 - \mu_2 \neq 0.$$

Samples of $n_1 = n_2 = 25$ tablets were drawn from each site, resulting in the following data:

Site 1		Site 2	
91.28	89.74	89.35	83.33
92.83	92.24	86.51	87.61
89.35	92.59	89.04	88.20
91.90	84.21	91.82	92.78
82.85	89.36	93.02	86.35
94.83	90.96	88.32	93.84
89.93	92.85	88.76	91.20
89.00	89.39	89.26	93.44
84.62	89.82	90.36	86.77
86.96	89.91	87.16	83.77
88.32	92.16	91.74	93.19
91.17	88.67	86.12	81.79
83.86		92.10	

Use the data to test the hypotheses at the $\alpha = 0.05$ significance level.

Answer: We first read the data into R with the following code:

```
site1 <- c( 91.28,92.83,89.35,91.90,82.85,94.83,
            89.93,89.00,84.62,86.96,88.32,91.17,
            83.86,89.74,92.24,92.59,84.21,89.36,
            90.96,92.85,89.39,89.82,89.91,92.16,
            88.67)

site2 <- c( 89.35,86.51,89.04,91.82,93.02,88.32,
            88.76,89.26,90.36,87.16,91.74,86.12,
            92.10,83.33,87.61,88.20,92.78,86.35,
            93.84,91.20,93.44,86.77,83.77,93.19,
            81.79)

n1 <- length(site1)
n2 <- length(site2)
```

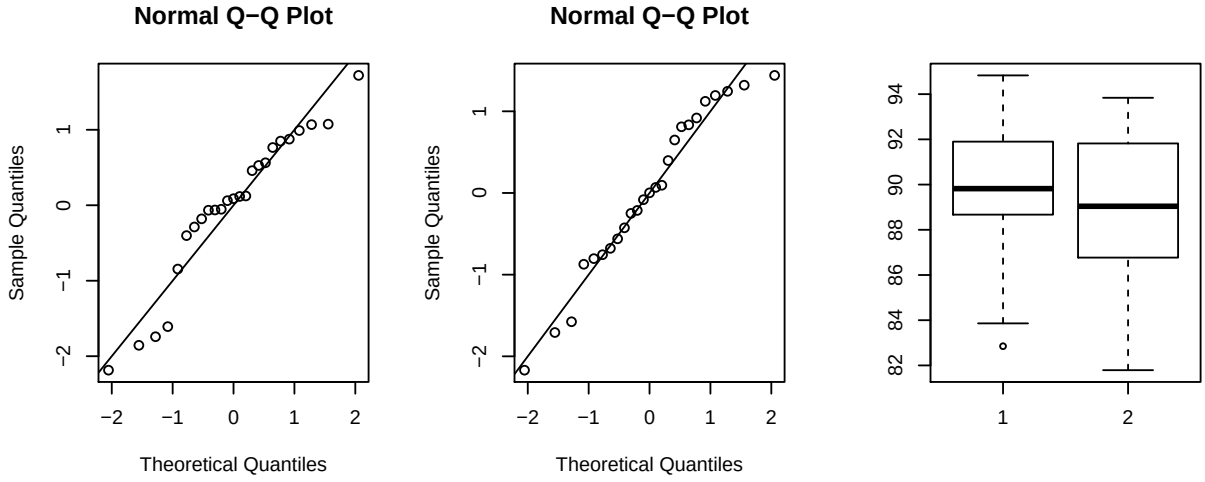
Now, since n_1 and n_2 are less than 30, we check whether the samples appear to have come from Normal populations. We can do this by looking at Normal quantile-quantile plots of both samples. In addition, we check to see whether we may assume equal variances for the two populations. We can do this by looking at side-by-side boxplots. The following R code produces these plots:

```
par(mfrow=c(1,3))

qqnorm(scale(site1))
abline(0,1)

qqnorm(scale(site2))
abline(0,1)

boxplot(site1,site2)
```



Based on the Normal quantile-quantile plots, it seems safe to assume that the samples have been drawn from Normal populations, as the points fall fairly close to the straight lines. In addition, based on the side-by-side boxplots we can probably safely assume that the variances σ_1^2 and σ_2^2 of the two populations are equal. Therefore, we will compute the test statistic

$$T_{\text{test}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S^2_{\text{pooled}}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

and reject the null hypotheses H_0: $\mu_1 - \mu_2 = 0$ if $|T_{\text{test}}| > t_{25+25-2,0.025} =$ `qt(0.975,48)` $=$ 2.010635. We can compute the test statistic and get the p-value using the R function `t.test()`. The command

```
t.test(site1,site2,var.equal=TRUE,alternative="two.sided")
```

returns the output

```
        Two Sample t-test

data:  site1 and site2
t = 0.57214, df = 48, p-value = 0.5699
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.304376  2.341976
sample estimates:
mean of x mean of y
  89.5520   89.0332
```

From which we pull the value of the test statistic $T_{\text{test}} = 0.57214$ and the corresponding p-value of 0.5699. The option `var.equal=TRUE` specifies that the variances σ_1^2 and σ_2^2 should be assumed equal and the option `alternative="two.sided"` specifies that the alternate hypothesis is the two sided one H_1: $\mu_1 - \mu_2 \neq 0$.

Since the absolute value of the test statistic 0.57214 does not exceed 2.010635, we fail to reject H_0 at the $\alpha = 0.05$ significance level. Equivalently, since the p-value 0.5699 is greater than $\alpha = 0.05$, we fail to reject H_0 at the $\alpha = 0.05$ significance level.

We could also perform the test of hypotheses by constructing a 95% confidence interval for $\mu_1 - \mu_2$ and checking whether it contains 0. We have

$$S^2_{\text{pooled}} = \texttt{sqrt((var(site1)*(n1-1) + var(site2)*(n2-1))/(n1+n2-2))} = 3.205903,$$

so that the 95% confidence interval is given by

$$89.5520 - 89.0332 \pm (2.010635)3.205903\sqrt{\left(\frac{1}{25} + \frac{1}{25}\right)} = (-1.304376, 2.341976).$$

Note that this same interval can be found in the output of the `t.test()` function. Since the 95% confidence interval contains 0, we fail to reject H_0 at the $\alpha = 0.05$ significance level.

If we did not think it was appropriate to assume equal variances, we would have specified `var.equal=FALSE` and executed the command

```
t.test(site1,site2,var.equal=FALSE,alternative="two.sided")
```

which produces the output

```
        Welch Two Sample t-test

data:  site1 and site2
t = 0.57214, df = 47.659, p-value = 0.5699
```

```
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.304712  2.342312
sample estimates:
mean of x mean of y
  89.5520   89.0332
```

Note that the degrees of freedom is now $\nu^* = 47.659$, which we could compute ourselves using the sample variances

$$S_1^2 = \texttt{var(site1)} = 9.4089 \quad \text{and} \quad S_2^2 = \texttt{var(site2)} = 11.14672$$

as

$$\nu^* = \left(\frac{9.4089}{25} + \frac{11.14672}{25}\right)^2 \left[\frac{(9.4089/25)^2}{25-1} + \frac{(11.14672/25)^2}{25-1}\right]^{-1} = 47.65936.$$

We see that the p-value is the same out to 4 decimal places (the degrees of freedom changed from 48 to 47.65936, which is a very small change). In addition, we see that the test statistic $T_{\text{test}} = 0.57214$ did not change. The reason for this is that the sample sizes are equal; when the sample sizes are equal, such that $n_1 = n_2 = n$, say, then we have

$$S^2_{\text{pooled}}\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = \frac{(n-1)(S_1^2+S_2^2)}{2(n-1)}\frac{2}{n} = \frac{S_1^2+S_2^2}{n}$$

as well as

$$\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} = \frac{S_1^2+S_2^2}{n}.$$

Therefore if $n_1 = n_2 = n$ we have

$$\underbrace{\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S^2_{\text{pooled}}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}}_{T_{\text{test}} \text{ under } \sigma_1^2 = \sigma_2^2} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2+S_2^2}{n}}} = \underbrace{\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)}}}_{T_{\text{test}} \text{ under } \sigma_1^2 \neq \sigma_2^2},$$

so that the equal-variances and unequal-variances test statistics will be the same when the sample sizes are equal. The degrees of freedom of the t-distribution from which we take the critical value will, however, change depending on whether we assume equal variances or not.

Concerning proportions

To compare the proportions p_1 and p_2, we will compare the sample proportions $\hat{p}_1$ and $\hat{p}_2$, specifically considering their difference $\hat{p}_1 - \hat{p}_2$. The following result about the sampling distribution of $\hat{p}_1 - \hat{p}_2$ will be our starting point for constructing confidence intervals for $p_1 - p_2$ and for building a test statistic for hypotheses concerning $p_1 - p_2$.

Sampling distribution result: Difference in sample proportions

Let $\hat{p}_1$ and $\hat{p}_2$ be the proportions of successes in two samples of sizes n_1 and n_2 from populations having success proportions p_1 and p_2, respectively. Then

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$ has a distribution more and more like the $\text{Normal}(0,1)$ dist.

for larger and larger sample sizes n_1 and n_2.

Rule of thumb 3. *If*

$$\min\{n_1\hat{p}_1, n_1(1-\hat{p}_1)\} \geq 15 \quad \textit{and} \quad \min\{n_2\hat{p}_2, n_2(1-\hat{p}_2)\} \geq 15 \tag{1}$$

is satisfied, we will assume that the sampling distribution of

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

is well-approximated by the $\text{Normal}(0,1)$ *distribution.*

Note that when we replace in the denominator the true proportions with the sample proportions we require larger sample sizes for assuming Normality.

The above gives us that

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

is a $(1-\alpha)\times 100\%$ confidence interval for $p_1 - p_2$, provided the condition in (1) is satisfied.

We now wish to build a test statistic for hypotheses of the form

$$\begin{array}{lll} H_0\colon\ p_1 - p_2 \geq 0 & \text{or} & H_0\colon\ p_1 - p_2 = 0 \quad \text{or} \quad H_0\colon\ p_1 - p_2 \leq 0 \\ H_1\colon\ p_1 - p_2 < 0 & & H_1\colon\ p_1 - p_2 \neq 0 \qquad\quad H_1\colon\ p_1 - p_2 > 0. \end{array}$$

Keep in mind that we always construct test statistics under the assumption that H_0 is true at the value which separates the null from the alternative. So in this case we will construct our test statistic based on the assumption that $p_1 - p_2 = 0$. Now, $p_1 - p_2 = 0$ means that $p_1 = p_2$, and it will be convenient for us to give a name to the common value of p_1 and p_2 under H_0. Let us call it p_0.

If we were assuming that the two populations had the same proportion p_0 of successes, which we are doing for the sake of building a test statistic, then a reasonable estimator for p_0 would be

$$\hat{p}_0 = \frac{\#\{\text{successes in sample one}\} + \#\{\text{successes in sample two}\}}{n_1 + n_2},$$

which is the total number of successes in both samples divided by the sum of the sample sizes. Then we find that a suitable test statistic is given by

$$Z_{\text{test}} = \frac{\hat{p}_1 - \hat{p}_2 - (0)}{\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_1} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_2}}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

The next result give the sampling distribution of this quantity when $p_1 - p_2 = 0$.

Sampling distribution result: Difference in sample proportions under null hypothesis

If $p_1 = p_2 = p_0$, then

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$ has a distribution more and more like the $\text{Normal}(0,1)$ dist

for larger and larger sample sizes n_1 and n_2.

Rule of thumb 4. *If*

$$\min\{n_1\hat{p}_0, n_1(1-\hat{p}_0)\} \geq 5 \quad \textit{and} \quad \min\{n_2\hat{p}_0, n_2(1-\hat{p}_0)\} \geq 5 \tag{2}$$

is satisfied, we will assume that the sampling distribution of

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

is well-approximated by the $\text{Normal}(0,1)$ *distribution.*

According to sampling distribution result , we can get a p-value associated with Z_{test} according to

$$p\text{-value} \overset{\text{approx}}{=} \begin{cases} P(Z < Z_{\text{test}}) & \text{if } H_1\text{: } p_1 - p_2 < 0 \\ 2P(Z > |Z_{\text{test}}|) & \text{if } H_1\text{: } p_1 - p_2 \neq 0 \\ P(Z > Z_{\text{test}}) & \text{if } H_1\text{: } p_1 - p_2 > 0 \end{cases},$$

where Z is a standard Normal random variable, provided that condition in (2) is satisfied.

Exercise. This exercise is inspired by the article [3], which was a survey study of 15-34 year-olds asking about the use of JUUL e-cigarettes. Suppose 6.0% of 1,000 randomly sampled 15-17 year-olds are found to have used JUUL within the last month and 3.0% of 1,000 randomly sampled 25-34 year-olds are found to have used JUUL within the last month (Disclaimer: These are not the actual numbers from the study, as the data from the study are not publicly available ☹. These numbers are loosely based on summary statistics described in the abstract of the paper).

1. Suppose you are interested in testing whether the proportion of JUUL users is greater in the younger age group. What are the hypotheses of interest?

   Answer: Letting p_1 be the proportion of JUUL users among 15-17 year-olds and p_2 be the proportion of JUUL users among 25-34 year-olds. Then we are interested in the hypotheses

   $$H_0\text{: } p_1 \leq p_2 \text{ versus } H_1\text{: } p_1 > p_2.$$

2. To what decision at the $\alpha = 0.01$ significance level does the data lead?

   Answer: We have $\hat{p}_1 = 0.06$ and $\hat{p}_2 = 0.03$. We compute $\hat{p}_0$ as

   $$\hat{p}_0 = \frac{1{,}000(0.06) + 1{,}000(0.03)}{2{,}000} = 0.045.$$

   The test statistic is thus

   $$Z_{\text{test}} = \frac{0.06 - 0.03}{\sqrt{0.045(1-0.045)\left(\frac{1}{1{,}000} + \frac{1}{1{,}000}\right)}} = 3.235924.$$

   The p-value is the area under the Normal$(0, 1)$ pdf to the right of this number. The p-value is $P(Z > 3.235924) =$ `1- pnorm(3.235924)` $= 0.0006$. Based on this data, we would reject the null hypotheses at $\alpha = 0.01$ (actually you would reject H_0 for any choices of α which are greater than 0.0006) and conclude that the proportion of JUUL users in the 15-17 year-old age group is greater than the proportion of JUUL users in the 25-34 year-old age group.

References

[1] Robert J MacG Dawson. The "unusual episode" data revisited. *Journal of Statistics Education*, 3(3), 1995.

[2] J.T. McClave and T.T. Sincich. *Statistics*. Pearson Education, 2016.

[3] Donna M Vallone, Morgane Bennett, Haijun Xiao, Lindsay Pitzer, and Elizabeth C Hair. Prevalence and correlates of juul use among a national sample of youth and young adults. *Tobacco Control*, 2018.