1. Testing hypotheses about populations proportions.
2. Comparing two population means (PNC 6.8)
Testing Hypotheses about a population proportion (PNC 6.7)

Examples of where this might be interesting

1. Is the proportion of defective products too high?
2. Is the cancer rate higher for people who live in a specific location than the rate in the general population?
3. Based on the proportion of offspring that have a specific trait, can we conclude anything about the genetic makeup of the parents?

1. Scientific Hypothesis
2. Statistical Model
   - We observe $Y$, the number out of the $n$ units in our sample that have some specified property.
   - The proportion of units in the population that have the specified property is some (unknown) value $p$ between 0 and 1.
   - Therefore, $Y \sim bin(n, p)$. That is, $Y$ follows a binomial distribution.
   - Based on what we learned about the binomial distribution on Chapter 4 (MA 2611) the mean of this population is $\mu = np$ and the variance is $\sigma^2 = np(1 - p)$.

3. Statistical hypotheses
   Depending on the scientific hypothesis, the statistical hypothesis will be one of the following, where $p_0$ is a specific proportion (i.e., a number between 0 and 1).

   \[
   H_0 : p = p_0 \quad H_0 : p = p_0 \quad H_0 : p = p_0 \\
   H_a : p < p_0 \quad H_a : p \neq p_0 \quad H_a : p > p_0
   \]
4. **Test statistic**

When $H_0$ is true $Y \sim bin(n, p_0)$, so our test statistic is defined

$$z^* = \frac{Y - \mu_0}{\sigma_0} = \frac{Y - np_0}{\sqrt{np_0(1 - p_0)}}$$

You will also see it written with the numerator in terms of the sample and population proportion.

$$z^* = \frac{(Y - np_0)/n}{\sqrt{np_0(1 - p_0)/n}} = \frac{Y/n - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

where $\hat{p}$ is the proportion of the $n$ units in our sample that have the specific property of interest.

5. **Sampling distribution**

When $H_0$ is true, $Y > 5$ and $(n - Y) > 5$ then $z^* \approx N(0, 1)$.

**Approximate?**

- We used the normal approximation to the binomial distribution.
- See Chapter 4 of the text for further discussion of this approximation.
- If $Y \leq 5$ or $(n - Y) \leq 5$ this approximation is not very good.
- In these cases we would need to use an alternate method. The “exact test” for these cases is discussed on pp. 311-313 of the text. We will not discuss the exact test in class.
6. **Drawing a conclusion about** $H_0$

To see whether or not we reject $H_0$ we can use one of the two equivalent methods.

(a) The $p$-value approach

i. Find the $p$-value just like you did when you tested the hypothesis about the population mean using $z^*$.

ii. If the $p$-value is less than $\alpha$, reject $H_0$. Otherwise do not reject $H_0$.

(b) The critical value approach

i. Find the $z_{crit}$ (in Table A.4) just like you did when you tested the hypothesis about the population mean using $t^*$. The $\text{N}(0,1)$ distribution is like a $t$ distribution with $\infty$ degrees of freedom.

ii. A. For $H_a : p < p_0$, $z_{crit} = -t_{\infty,1-\alpha}$ and if $z^* < z_{crit}$ reject $H_0 : p = p_0$ in favor of $H_a$. Otherwise do not reject $H_0$.

B. For $H_a : p > p_0$, $z_{crit} = t_{\infty,1-\alpha}$ and if $z^* > z_{crit}$ reject $H_0 : p = p_0$ in favor of $H_a$. Otherwise do not reject $H_0$.

C. For $H_a : p \neq p_0$, $z_{crit} = t_{\infty,1-\alpha/2}$ and if $|z^*| > z_{crit}$ reject $H_0 : p = p_0$ in favor of $H_a$. Otherwise do not reject $H_0$.
Example:
Studies have shown that the proportion of left-handed males in the general population is $\frac{1}{5}$. There are also those that suggest that left-handed people are more intelligent on average. Assuming that WPI students (for the most part) would be considered “intelligent,” we ask a random sample of 100 male WPI students whether or not they are left-handed. We find that 15 out of the 100 are left-handed.

Scientific hypothesis

Statistical model

Statistical hypotheses

Test statistic
Conclusion

1. $p$-value approach

2. Critical value approach
Comparing two population means (PNC 6.8)

Examples of where this might be interesting:

1. Does a drug affect blood pressure levels?
2. Is there a bias in the SAT that favors males?

Scientific hypothesis

- There are a number of variations on the test for comparing the means of two populations.
- We cannot tell which test to use based on the scientific hypothesis alone.
- The choice of which test to use in a particular situation depends on the experimental design and the assumptions of the statistical model.
**Paired comparisons**

**Statistical model**

- \((Y_{1,1}, Y_{2,1}), (Y_{1,2}, Y_{2,2}), \ldots, (Y_{1,n}, Y_{2,n})\) are pairs of observations made on the same experimental unit. (e.g., Measurements of cholesterol before and after a drug treatment on the same patient.)

- We observe the difference between \(Y_{1,i}\) and \(Y_{2,i}\), \(D_i = Y_{1,i} - Y_{2,i}\) for each \(i = 1, \ldots, n\).

- We assume that these differences are a random sample from a normal distribution: \(D_1, D_2, \ldots, D_n \sim N(\mu_D, \sigma_D^2)\)

**Statistical hypothesis**

\[
H_0 : \mu_D = \delta_0 \quad H_0 : \mu_D = \delta_0 \quad H_0 : \mu_D = \delta_0 \\
H_a : \mu_D < \delta_0 \quad H_a : \mu_D \neq \delta_0 \quad H_a : \mu_D > \delta_0
\]

For most scientific hypotheses \(\delta_0 = 0\). This corresponds to null hypothesis of no change.

**Test statistic**

\[
t^* = \frac{\bar{D} - \delta_0}{s_D / \sqrt{n}}
\]

where \(s_D\) is the usual estimate of \(\sigma_D\) based on \(D_1, D_2, \ldots D_n\).

\[
s_D = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (D_i - \bar{D})^2}
\]

**Sampling distribution**

- If \(H_0\) is true, \(t^*\) follows a \(t_{n-1}\) distribution.

- Based on this sampling distribution we can make a conclusion about \(H_0\) using either the \(p\)-value or the critical value approach.
Comparing the means of two independent populations

Statistical model

\[ Y_{1,1}, Y_{1,2}, \ldots, Y_{1,n_1} \sim N(\mu_1, \sigma_1^2) \]
\[ Y_{2,1}, Y_{2,2}, \ldots, Y_{2,n_2} \sim N(\mu_2, \sigma_2^2) \]

Statistical hypothesis

\[ H_0 : \mu_1 - \mu_2 = \delta_0 \]
\[ H_a : \mu_1 - \mu_2 < \delta_0 \]
\[ H_0 : \mu_1 - \mu_2 = \delta_0 \]
\[ H_a : \mu_1 - \mu_2 = \delta_0 \]
\[ H_0 : \mu_1 - \mu_2 = \delta_0 \]
\[ H_a : \mu_1 - \mu_2 > \delta_0 \]

For most scientific hypotheses \( \delta_0 = 0 \). This corresponds to null hypothesis of no difference.

Test statistic

\[ t^* = \frac{(\bar{Y}_1 - \bar{Y}_2) - \delta_0}{s(\bar{Y}_1 - \bar{Y}_2)} \]

where \( s(\bar{Y}_1 - \bar{Y}_2) \) is the standard error of \( (\bar{Y}_1 - \bar{Y}_2) \).

- The standard error is another name for the standard deviation of the sampling distribution.
- Thus, \( s(\bar{Y}_1 - \bar{Y}_2) \) will depend on any assumptions we make in the statistical model for \( Y_{1,1}, \ldots, Y_{1,n_1} \) and \( Y_{2,1}, \ldots, Y_{2,n_2} \).
- This will also affect the sampling distribution of \( t^* \) as well as any conclusions we make about \( H_0 \).
Computation of $s(\bar{Y}_1 - \bar{Y}_2)$ and the sampling distribution of $t^*$

Recall the statistical model:

$$Y_{1,1}, Y_{1,2}, \ldots, Y_{1,n_1} \sim N(\mu_1, \sigma_1^2)$$

$$Y_{2,1}, Y_{2,2}, \ldots, Y_{2,n_2} \sim N(\mu_2, \sigma_2^2)$$

1. With no further assumptions

$$s(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

where

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (Y_{1,i} - \bar{Y}_1)^2$$

$$S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_{2,i} - \bar{Y}_2)^2$$

and

$$t^* = \frac{(\bar{Y}_1 - \bar{Y}_2) - \delta_0}{s(\bar{Y}_1 - \bar{Y}_2)}$$

follows a $t_\nu$ distribution where $\nu$ is the largest integer less than or equal to

$$\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2 \left(\frac{S_1^2/n_1}{n_1 - 1} + \frac{S_2^2/n_2}{n_2 - 1}\right)$$

2. However, if it is reasonable to assume that $\sigma_1^2 = \sigma_2^2$ then

$$s(\bar{Y}_1 - \bar{Y}_2) = \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

and $t^*$ follows a $t_{n_1 + n_2 - 2}$ distribution.