# STAT509: Inference on Two Populations

Peijie Hou

University of South Carolina

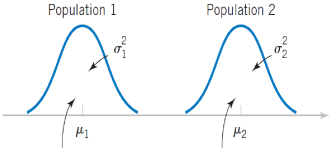
October 25, 2014

# Motivation of Inference on Two Samples

- Until now we have been mainly interested in a single parameter, either $\mu$ or $p$, of a single population.
- In practice, it is very common to compare the same characteristic (mean, proportion, variance) from two different distributions. For example, we may wish to compare
  - the **mean** starting salaries of male and female engineers (compare $\mu_1$ and $\mu_2$; independent vs. dependent samples)
  - the **proportion** of scrap produced from two manufacturing processes (compare $p_1$ and $p_2$)
  - the the **variance** of sound levels from two indoor swimming pool designs (compare $\sigma_1^2$ and $\sigma_2^2$).

# Inference on $\mu_1$ and $\mu_2$, assume $\sigma_1^2$ and $\sigma_2^2$ Known

- We want to compare the population means $\mu_1$ and $\mu_2$ of two normal populations.




- **Question:** Are $\mu_1$ and $\mu_2$ significantly different from each other?
- Suppose that we have two independent samples of size $n_1$ and $n_2$, respectively:

$$\textbf{Sample 1:} Y_{11}, Y_{12}, \ldots, Y_{1n_1} \sim \mathcal{N}(\mu_1, \sigma_1^2)$$

$$\textbf{Sample 2:} Y_{21}, Y_{22}, \ldots, Y_{2n_2} \sim \mathcal{N}(\mu_2, \sigma_2^2).$$

Assume $\sigma_1^2$ and $\sigma_2^2$ are known.

- A logical point estimator of $\mu_1 - \mu_2$ is the difference in sample mean $\overline{Y}_1 - \overline{Y}_2$.
- We already known that

$$\overline{Y}_1 \sim \mathcal{N}(\mu_1, \sigma_1^2/n_1) \text{ and } \overline{Y}_2 \sim \mathcal{N}(\mu_2, \sigma_2^2/n_2).$$

- By property of normal distribution, $\overline{Y}_1 - \overline{Y}_2$ is also normally distributed.
- We need to calculate mean and variance of $\overline{Y}_1 - \overline{Y}_2$.

$$E(\overline{Y}_1 - \overline{Y}_2) =$$

and

$$\mathrm{Var}\,(\overline{Y}_1 - \overline{Y}_2) =$$

- In summary, we have

$$\overline{Y}_1 - \overline{Y}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

# CI of $\mu_1 - \mu_2$, $\sigma_1^2$ and $\sigma_2^2$ Known

A $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is

$$\overline{y}_1 - \overline{y}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \overline{y}_1 - \overline{y}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Proof: see confidence interval for population proportion

# Hypothesis Testings on $\mu_1 - \mu_2$, Variances Known

- Recall we have

$$\overline{Y}_1 - \overline{Y}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

- **Step 1:** The null and alternative hypothesis

$$\begin{aligned} H_0 &: \mu_1 - \mu_2 = 0 \\ H_a &: \mu_1 - \mu_2 <, > \text{ or } \neq 0 \end{aligned}$$

- **Step 2:** The test statistic is

$$z_0 = \frac{\overline{y}_1 - \overline{y}_2(-0)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

# Hypothesis Testings on $\mu_1 - \mu_2$, Variances Known

- **Step 3:** $p$-value
  - For test $H_a : \mu_1 < \mu_2$, $p$-value$= P(Z < z_0)$;
  - For test $H_a : \mu_1 > \mu_2$, $p$-value$= P(Z > z_0)$;
  - For test $H_a : \mu_1 \neq \mu_2$, $p$-value$= 2P(Z < -|z_0|)$.

# Example: Drying Time

A product developer is interested in reducing the drying time of a primer paint. Two formulations of the paint are tested; formulation 1 is the standard chemistry, and formulation 2 has a new drying ingredient that should reduce the drying time. From experience, it is known that the standard deviation of drying time is 8 minutes, and this inherent variability should be unaffected by the addition of the new ingredient. Ten specimens are painted with formulation 1, and another 10 specimens are painted with formulation 2; the 20 specimens are painted in random order. The two sample average drying times are $\overline{y}_1 = 121$ minutes and $\overline{y}_2 = 112$ minutes, respectively.

What conclusion can the product developer draw about the effectiveness of the new ingredient, using $\alpha = 0.05$?

# Example: Drying Time

Solution:

- Step 1: state hypothesis
- Step 2: Test statistic
- Step 3: $p$-value
- Step 4: Decision and conclusion

# Inference on $\mu_1$ and $\mu_2$, assume unknown $\sigma_1^2$ and $\sigma_2^2$

- The construction of confidence intervals and hypothesis testings depend on the values of $\sigma_1^2$ and $\sigma_2^2$.
  1. $\sigma_1^2 = \sigma_2^2$ (equal variance case),
  2. $\sigma_1^2 \neq \sigma_2^2$ (unequal variance case)
- We first consider the case $\sigma_1^2 = \sigma_2^2$.
- Just like inference for single proportion, single mean, and single variance, we need a sampling distribution involving $\mu_1 - \mu_2$. It can be shown that

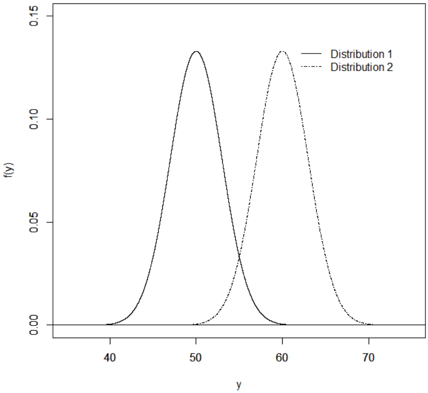
$$T = \frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2(\frac{1}{n_1} + \frac{1}{n_2})}} \sim t(n_1 + n_2 - 2),$$

where

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

# Inference on Two Means: $\sigma_1^2 = \sigma_2^2$

Assuming the variances of the two distributions are the same, we want to test whether two populations (means) are the same or not.



# CI for $\mu_1 - \mu_2$: assume $\sigma_1^2 = \sigma_2^2$

- A $(1-\alpha)100\%$ confidence interval of $\mu_1 - \mu_2$ is given by

$$\underbrace{(\overline{y}_1 - \overline{y}_2)}_{\text{pt. est.}} \pm \underbrace{t_{\alpha/2, n_1+n_2-2}}_{\text{quantile}} \underbrace{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}_{\text{standard error}}.$$

- **Interpretation**: we are $(1-\alpha)100\%$ confident that the population mean difference $\mu_1 - \mu_2$ is in this interval (in the context of the question).

# Hypothesis Testing on $\mu_1 - \mu_2$, assume $\sigma_1^2 = \sigma_2^2$

- We want to test $H_0 : \mu_1 - \mu_2 = 0$ against three types of alternative.
- The test statistic is

$$t_0 = \frac{\overline{y}_1 - \overline{y}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

- For test $H_a : \mu_1 \neq \mu_2$,, the $p$-value is $2P(T_{n_1+n_2-2} < -|t_0|)$;
- For test $H_a : \mu_1 < \mu_2$,, the $p$-value is $P(T_{n_1+n_2-2} < t_0)$;
- For test $H_a : \mu_1 > \mu_2$, the $p$-value is $P(T_{n_1+n_2-2} > t_0)$.

# Example: Fish Weights

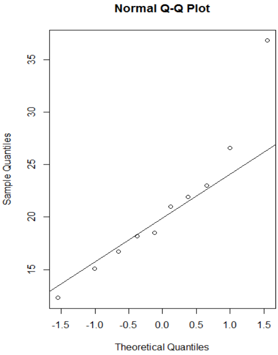
- In the vicinity of a nuclear power plant, environmental engineers from the EPA would like to determine if there is a difference between the mean weight in fish (of the same species) from two locations. Independent samples are taken from each location and the following weights (in ounces) are observed:

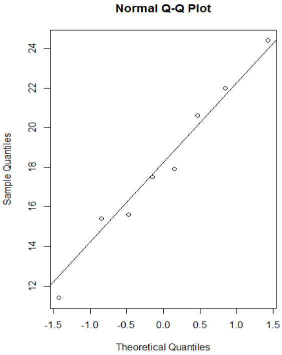
| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Location 1: | 21.9 | 18.5 | 12.3 | 16.7 | 21.0 | 15.1 | 18.2 | 23.0 | 36.8 | 26.6 |
| Location 2: | 22.0 | 20.6 | 15.4 | 17.9 | 24.4 | 15.6 | 11.4 | 17.5 | | |

- **Question:** Let $\mu_i$ denotes the mean weight of fish in location $i$, $i = 1, 2$. Are $\mu_1$ and $\mu_2$ significantly different from each other?

# Check Assumptions

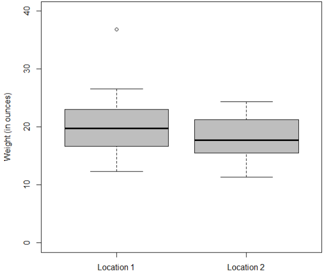
We start by checking normality assumption through Q-Q plots. The points are approximately around a straight line, which indicates the normality assumption is reasonable for both populations.



# Check Assumptions Cont'd

In order to visually assess the equal variance assumption, we use boxplots to display the data in each sample.




```
loc.1 = c(21.9,18.5,12.3,16.7,21.0,15.1,18.2,23.0,36.8,26.6)
loc.2 = c(22.0,20.6,15.4,17.9,24.4,15.6,11.4,17.5)
# Create side by side boxplots
boxplot(loc.1,loc.2,xlab="",names=c("Location 1","Location 2"),
ylab="Weight (in ounces)",ylim=c(0,40),col="grey")
```

# Check Assumptions Cont'd

- To be formal, we can also perform a hypothesis test of

$$H_0 : \sigma_1^2/\sigma_2^2 = 1,$$

$$H_a : \sigma_1^2/\sigma_2^2 \neq 1.$$

- The R output is

```
loc.1 = c(21.9,18.5,12.3,16.7,21.0,15.1,18.2,23.0,36.8,26.6)
loc.2 = c(22.0,20.6,15.4,17.9,24.4,15.6,11.4,17.5)
var.test(loc.1, loc.2 )

        F test to compare two variances

data:  loc.1 and loc.2
F = 2.7803, num df = 9, denom df = 7, p-value = 0.1914
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  0.5764409 11.6690466
sample estimates:
ratio of variances
          2.780299
```

# Example: Fish Weights

- Recall: A $(1-\alpha)100\%$ confidence interval of $\mu_1 - \mu_2$ is given by

$$\underbrace{(\bar{Y}_1 - \bar{Y}_2)}_{\text{pt. est.}} \pm \underbrace{t_{n_1+n_2-2,\alpha/2}}_{\text{quantile}} \underbrace{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}}_{\text{standard error}}.$$

- The sample summary statistic is

| Variable | $n$ | mean | variance |
|---|---|---|---|
| Location 1 | 10 | 21.01 | 47.65 |

| Variable | $n$ | mean | variance |
|---|---|---|---|
| Location 2 | 8 | 18.1 | 17.14 |

$t_{0.025,16} = 2.12$.

# Example: Fish Weights Cont'd

- The R output is

```
> t.test(loc.1,loc.2,conf.level=0.90,var.equal=TRUE)

        Two Sample t-test

data:  loc.1 and loc.2
t = 1.0474, df = 16, p-value = 0.3105
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
 -1.940438  7.760438
sample estimates:
mean of x mean of y
    21.01     18.10
```

- **Interpretation:** We are 90% confident that the mean weight difference in fish for location 1 and 2 is between -1.94 and 7.76 oz.
- Note that this interval includes "0". By the relationship between hypothesis test and confidence interval, we fail to reject $H_0 : \mu_1 - \mu_2 = 0$ against $H_a : \mu_1 - \mu_2 \neq 0$. at $\alpha = 0.1$ level.
- Therefore, we do not have sufficient evidence that the population mean fish weights $\mu_1$ and $\mu_2$ are different at 0.1 level of significance.

# Comments on Confidence Interval

1. We should only use this interval if there is strong evidence that the population variances $\sigma_1^2$ and $\sigma_2^2$ are equal (or at least close). Otherwise, we should use a different interval (coming up).

2. The two sample $t$ interval (and the unequal variance version coming up) is robust to normality departures. This means that we can feel comfortable with the interval even if the underlying population distributions are not perfectly normal.

# Inference on Two Means: $\sigma_1^2 \neq \sigma_2^2$

- When $\sigma_1^2 \neq \sigma_2^2$, the construction of a $(1-\alpha)100\%$ is no longer based on the exact $t$ distribution.
- We need to approximate the degrees of freedom of a $t$ distribution, and write an **approxiamte** confidence interval.
- An approximate $(1-\alpha)100\%$ confidence interval $\mu_1 - \mu_2$ is given by

$$(\overline{y}_1 - \overline{y}_2) \pm t_{\alpha/2,\nu}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

where

$$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}.$$

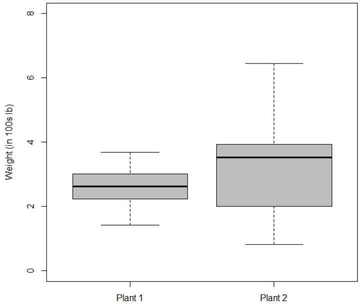
# Example: Recycling Project

You are part of a recycling project that is examining how much paper is being discarded (not recycled) by employees at two large plants. These data are obtained on the amount of white paper thrown out per year by employees (data are in hundreds of pounds). Samples of employees at each plant were randomly selected.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Plant 1: | 3.01 | 2.58 | 3.04 | 1.75 | 2.87 | 2.75 | 2.51 | 2.93 | 2.85 | 3.09 |
| | 1.43 | 3.36 | 3.18 | 2.74 | 2.25 | 1.95 | 3.68 | 2.29 | 1.86 | 2.63 |
| | 2.83 | 2.04 | 2.23 | 1.92 | 3.02 | | | | | |
| Plant 2: | 3.79 | 2.08 | 3.66 | 1.53 | 4.07 | 4.31 | 2.62 | 4.52 | 3.80 | 5.30 |
| | 3.41 | 0.82 | 3.03 | 1.95 | 6.45 | 1.86 | 1.87 | 3.78 | 2.74 | 3.81 |

# Example: Recycling Project Cont'd

Let us look at the side-by-side plot to check the variances:




It is quite obvious that $\sigma_1^2 \neq \sigma_2^2$. A formal $F$ test can be conducted to confirm it.

# Example: Recycling Project Cont'd

- A 90% approximate (unequal variance) confidence interval for $\mu_1 - \mu_2$ is given by

```
> plant.1 = c(3.01,2.58,3.04,1.75,2.87,2.57,2.51,2.93,2.85,3.09,
+ 1.43,3.36,3.18,2.74,2.25,1.95,3.68,2.29,1.86,2.63,
+ 2.83,2.04,2.23,1.92,3.02)
> plant.2 = c(3.79,2.08,3.66,1.53,4.07,4.31,2.62,4.52,3.80,5.30,
+ 3.41,0.82,3.03,1.95,6.45,1.86,1.87,3.78,2.74,3.81)
> # Calculate t interval directly
> t.test(plant.1,plant.2,conf.level=0.95,var.equal=FALSE)

        Welch Two Sample t-test

data:  plant.1 and plant.2
t = -2.1037, df = 23.972, p-value = 0.04608
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.35825799 -0.01294201
sample estimates:
mean of x mean of y
   2.5844    3.2700
```

# Example: Recycling Project Cont'd

We can also test for $H_a : \mu_1 - \mu_2 < 0$ and $H_a : \mu_1 - \mu_2 > 0$ by specifying "alternative".

```
> t.test(plant.1,plant.2,conf.level=0.95,var.equal=FALSE,alternative="less")

        Welch Two Sample t-test

data:  plant.1 and plant.2
t = -2.1037, df = 23.972, p-value = 0.02304
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.1280042
sample estimates:
mean of x mean of y
   2.5844    3.2700

> t.test(plant.1,plant.2,conf.level=0.95,var.equal=FALSE,alternative="greater")

        Welch Two Sample t-test

data:  plant.1 and plant.2
t = -2.1037, df = 23.972, p-value = 0.977
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -1.243196       Inf
sample estimates:
mean of x mean of y
   2.5844    3.2700
```

# Two Population Means: Dependent Sample

- Suppose the populations are not independent.
- For example:
  1. Weights before and after a given diet.
  2. Measurements obtained by two different instruments.
  3. Comparison of two medical treatments.
- **Basic idea:** In the dependent sample problem, the data is presented in *matched pairs*. In general, we apply two treatments (e.g., with/without diet, measurements by instrument 1 and instrument 2, etc.) on the experiment units (e.g., individuals, products, etc.). We want to compare the two treatment effects by converting the matched pairs to one sample problem.
- The variation in the two independent sample contains variation among subjects and within subjects, while the source variation of matched pair only comes from within subjects.
- When you remove extra variability, this enables you to do a better job at comparing the two experimental conditions.

# Example: Creatine

- Creatine is an organic acid that helps to supply energy to cells in the body, primarily muscle. Because of this, it is commonly used by those who are weight training to gain muscle mass. Suppose that we are designing an experiment involving USC male undergraduates who exercise/lift weights regularly.
- **Question:** does creatine really work?

| Student $j$ | Creatine BP | Control BP | Difference ($D_j = Y_{1j} - Y_{2j}$) |
|---|---|---|---|
| 1 | 230 | 200 | 30 |
| 2 | 140 | 155 | -15 |
| 3 | 215 | 205 | 10 |
| 4 | 190 | 190 | 0 |
| 5 | 200 | 170 | 30 |
| 6 | 230 | 225 | 5 |
| 7 | 220 | 200 | 20 |
| 8 | 255 | 260 | -5 |
| 9 | 220 | 240 | -20 |
| 10 | 200 | 195 | 5 |
| 11 | 90 | 110 | -20 |
| 12 | 130 | 105 | 25 |
| 13 | 255 | 230 | 25 |
| 14 | 80 | 85 | -5 |
| 15 | 265 | 255 | 10 |

# Implementation

- Data from matched pairs experiments are analyzed by examining the difference in responses of the two treatments.
- Specifically, compute $D_j = Y_{1j} - Y_{2j}$ for each individual $j = 1, 2, \ldots, n$.
- Using the formula for one sample mean, a $(1-\alpha)100\%$ C.I. for one-sample data $D_j$ is

$$\overline{Y}_D \pm t_{n-1,\alpha/2}\frac{S_D}{\sqrt{n}},$$

where $\overline{Y}_D$ and $S_D$ are sample mean and s.d. based on $d_1, \ldots, d_n$.

# Implementation Cont'd

- Hypothesis testing for test $H_0 : \mu_D = 0$ against three types of alternative, see inference on single mean with $\overline{Y}$ replaced by $\overline{Y}_D$, and $S$ replaced by $S_D$.

# Example: Creatine

- The R output for creatine data is

```
creatine = c(230,140,215,190,200,230,220,255,220,200,90,130,255,80,265)
control = c(200,155,205,190,170,225,200,260,240,195,110,105,230,85,255)
> t.test(creatine,control, conf.level=0.95,alternative="greater",paired=TRUE)

        Paired t-test

data:  creatine and control
t = 1.4207, df = 14, p-value = 0.08865
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -1.518439       Inf
sample estimates:
mean of the differences
               6.333333
```

- **Interpretation**: The $p$-value of the test is 0.08865, we will reject $H_0$ at $\alpha = 0.10$ level. This suggests that taking creating leads to a larger mean MBPW. However, the evidence is not that strong. (we will fail to reject if we use $\alpha = 0.05$)

# Inference on Two Population Proportions

- Just like we extend inference for one sample mean to two sample means, we also can extend our inference procedure for a single population proportion $p$ to **two populations**.
- Define
  $p_1$=population proportion of successes in Population 1,
  $p_2$=population proportion of successes in Population 2.
- For example, we can compare
  1. defective rate of water filters for two different suppliers
  2. the proportion of on-time payments for two classes of customers
  3. The proportion of exceedence of the localizer for two training programs.

# Point Estimator of $p_1 - p_2$

- Let $Y_i$ =number of "successes" in the $i^{\text{th}}$ sample out of $n_i$ individuals, it follows that $Y_i \sim \text{binomial}(n_i, p_i)$, for $i = 1, 2$.
- We know that the most efficient estimators for $p_1$ and $p_2$ are $\hat{p}_1 = \frac{Y_1}{n_1}$ and $\hat{p}_2 = \frac{Y_2}{n_2}$, respectively. A natural point estimator for $p_1 - p_2$ is given by

$$\hat{p}_1 - \hat{p}_2 = \frac{Y_1}{n_1} - \frac{Y_2}{n_2}.$$

# Sampling Distribution

Recall that we have following approximate sampling distribution of $\hat{p}_1$ and $\hat{p}_2$:

$$\hat{p}_1 \sim \mathcal{AN}\left(p_1, \frac{p_1(1-p_1)}{n_1}\right),$$

and

$$\hat{p}_2 \sim \mathcal{AN}\left(p_2, \frac{p_2(1-p_2)}{n_2}\right).$$

It can be shown that

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \sim \mathcal{AN}(0, 1).$$

# Two Population Proportions: Confidence interval

- To make the above sampling distribution valid, We need
  1. the two samples to be independent
  2. the sample size $n_1$ and $n_2$ to be "large", i.e., $n_i\hat{p}_i > 15$, and $n_i(1-\hat{p}_i) > 15$ for $i = 1, 2$.
- We can use above distributional result to conduct a hypothesis test or construct an approximate $(1-\alpha)100\%$ confidence interval.
- Note again the form of the interval:

$$\underbrace{\text{Point estimate}}_{\hat{p}_1-\hat{p}_2} \pm \underbrace{\text{quantile}}_{z_{\alpha/2}} \times \underbrace{\text{standard error}}_{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

# Two Population Proportions: Hypothesis testing

- We want to test $p_1 = p_2$, which is equivalent to test $H_0 : p_1 - p_2 = 0$ vs. $H_a : p_1 - p_2 \neq (> \textit{ or } <)0$.
- The most efficient estimator for $p_1 = p_2 (= p_0)$ is

$$\hat{p}_0 = \frac{Y_1 + Y_2}{n_1 + n_2}$$

with $\hat{q}_0 = 1 - \hat{p}_0$. This leads to

$$\begin{aligned} \mathrm{Var}\left(\hat{p}_1 - \hat{p}_2\right) &= \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2} \\ &= \frac{p_0 q_0}{n_1} + \frac{p_0 q_0}{n_2} \\ &= p_0 q_0 \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \end{aligned}$$

# Two Population Proportions: Hypothesis testing

- An estimator of $\mathrm{Var}(\hat{p}_1 - \hat{p}_2)$ is $\hat{p}\hat{q}(\frac{1}{n_1} + \frac{1}{n_2})$.
- The test statistic under the null is

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

- For test $H_a : p_1 \neq p_2$,, the $p$-value is $2P(Z < -|z_0|)$;
- For test $H_a : p_1 < p_2$,, the $p$-value is $P(Z < z_0)$;
- For test $H_a : p_1 > p_2$, the $p$-value is $P(Z > z_0)$.

# Example: Proportion of Exceedence

- Airplanes approaching the runway for landing are required to stay within the localizer (a certain distance left and right of the runway). When an airplane deviates from the localizer, it is sometimes referred to as an exceedence. Consider two airlines at a large airport. During a three-week period, airline 1 had 14 exceedences out of 156 flights and airline 2 had 11 exceedences out of 198 flights.
- We are interested in whether airline 1 has a significantly higher exceedence rate than airline 2 or not.

# Example: Proportion of Exceedence Cont'd

Solution:

You can also use R:

```
> prop.test(c(14,11),c(156,198),alternative="greater",correct=F)

        2-sample test for equality of proportions without continuity
        correction

data:  c(14, 11) out of c(156, 198)
X-squared = 1.5538, df = 1, p-value = 0.1063
alternative hypothesis: greater
95 percent confidence interval:
 -0.01200417  1.00000000
sample estimates:
    prop 1     prop 2
0.08974359 0.05555556
```