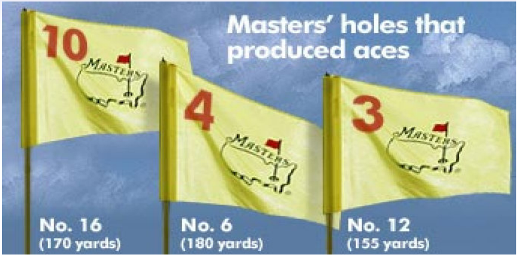
Trevor Immelman of South Africa was the last golfer to make a hole in one at The Masters, acing the 16th hole in 2005. Eighteen holes in one have been scored during Masters play.



By Kevin Greer and Ron Coddington, USA TODAY
Source: The Masters

Comparing Two Population Proportions

Comparing Two Population Proportions

Suppose there are two populations: population 1, in which a proportion p_1 have a certain characteristic, and population 2, in which a proportion p_2 have a certain (possibly different) characteristic. We will use a sample of size n_1 from population 1, and n_2 from population 2 to estimate the difference $p_1 - p_2$.

$\approx$

Specifically, if y_1 is the number having the population 1 characteristic in the n_1 items in sample 1, and if if y_2 is the number having the population 2 characteristic in the n_2 items in sample 2, then the sample proportion having the population 1 characteristic is $\hat{p}_1 = y_1/n_1$, and the sample proportion having the population 2 characteristic is $\hat{p}_2 = y_2/n_2$.

Comparing Two Population Proportions

Specifically, if y_1 is the number having the population 1 characteristic in the n_1 items in sample 1, and if if y_2 is the number having the population 2 characteristic in the n_2 items in sample 2, then the sample proportion having the population 1 characteristic is $\hat{p}_1 = y_1/n_1$, and the sample proportion having the population 2 characteristic is $\hat{p}_2 = y_2/n_2$.

A point estimator of $p_1 - p_2$ is $\hat{p}_1 - \hat{p}_2$.

$\approx$

Comparing Two Population Proportions

The standard error of $\hat{p}_1 - \hat{p}_2$ is

$$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$$

Further, for large n_1 and n_2, the Central Limit Theorem ensures that $\hat{p}_1 - \hat{p}_2$ has approximately a normal distribution, so

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

has approximately a $N(0,1)$ distribution.

$\approx$

Based on this, and on the fact that if n_1 and n_2 are large, then $\hat{p}_1$ and $\hat{p}_2$ are close to p_1 and p_2, respectively, an approximate level L confidence interval for $p_1 - p_2$ has endpoints

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

$\approx$

As for the one sample case, this large-sample interval does not work well when one or both sample sizes are small.

As for the one sample case, this large-sample interval does not work well when one or both sample sizes are small.

However, by "fudging" the sample proportions in much the same way as we did in the one sample case, we can get an approximate interval that works well for all sample sizes.

$\approx$

Comparing Two Population Proportions

Specifically, to compute the level L approximate score (or Agresti-Coull) interval, first compute the adjusted estimates of n_1 and n_2:

$$\tilde{n}_1 = n_1 + 0.5z^2_{(1+L)/2}, \ \tilde{n}_2 = n_2 + 0.5z^2_{(1+L)/2},$$

Comparing Two Population Proportions

Specifically, to compute the level L approximate score (or Agresti-Coull) interval, first compute the adjusted estimates of n_1 and n_2:

$$\tilde{n}_1 = n_1 + 0.5z^2_{(1+L)/2}, \;\; \tilde{n}_2 = n_2 + 0.5z^2_{(1+L)/2},$$

and then the adjusted estimates of p_1 and p_2:

$$\tilde{p}_1 = \frac{y_1 + 0.25z^2_{(1+L)/2}}{\tilde{n}_1}, \;\; \tilde{p}_2 = \frac{y_2 + 0.25z^2_{(1+L)/2}}{\tilde{n}_2}$$

Comparing Two Population Proportions

Specifically, to compute the level L approximate score (or Agresti-Coull) interval, first compute the adjusted estimates of n_1 and n_2:

$$\tilde{n}_1 = n_1 + 0.5z^2_{(1+L)/2}, \ \tilde{n}_2 = n_2 + 0.5z^2_{(1+L)/2},$$

and then the adjusted estimates of p_1 and p_2:

$$\tilde{p}_1 = \frac{y_1 + 0.25z^2_{(1+L)/2}}{\tilde{n}_1}, \ \tilde{p}_2 = \frac{y_2 + 0.25z^2_{(1+L)/2}}{\tilde{n}_2}$$

The approximate score interval for $p_1 - p_2$ is then given by the formula:

$$\tilde{p}_1 - \tilde{p}_2 \pm z_{(1+L)/2}\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{\tilde{n}_1} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{\tilde{n}_2}}$$

Example 4

In a recent survey on academic dishonesty 24 of the 200 female college students surveyed and 26 of the 100 male college students surveyed agreed or strongly agreed with the statement "Under some circumstances academic dishonesty is justified." With 95% confidence estimate the difference in the proportions p_f of all female and p_m of all male college students who agree or strongly agree with this statement.

$\approx$

Example 4

1. **The Scientific Goal:**

Example 4

1. **The Scientific Goal:** Estimate the difference in the proportions p_f of all female and p_m of all male college students who agree or strongly agree with the statement.
2. **The Statistical Model:**

Example 4

1. **The Scientific Goal:** Estimate the difference in the proportions p_f of all female and p_m of all male college students who agree or strongly agree with the statement.
2. **The Statistical Model:** Two independent binomials $b(200, p_f)$, $b(100, p_m)$.
3. **The Model Parameter(s) to Be Estimated:**

Example 4

1. **The Scientific Goal:** Estimate the difference in the proportions p_f of all female and p_m of all male college students who agree or strongly agree with the statement.
2. **The Statistical Model:** Two independent binomials $b(200, p_f),\ b(100, p_m)$.
3. **The Model Parameter(s) to Be Estimated:** $p_f - p_m$

$\approx$

Example 4

4. **Point and Interval Estimates:**
   a. Point estimate:

Example 4

4. **Point and Interval Estimates:**
   a. Point estimate: $\hat{p}_f - \hat{p}_m = \frac{24}{200} - \frac{26}{100} = -0.14$.
   b. Confidence interval:

Example 4

4. **Point and Interval Estimates:**

a. Point estimate: $\hat{p}_f - \hat{p}_m = \frac{24}{200} - \frac{26}{100} = -0.14$.

b. Confidence interval: Since $z_{0.975} = 1.96$, $y_f = 24$, $n_f = 200$, $y_m = 26$, and $n_m = 100$, the adjusted estimates of n_f and n_m are

$$\tilde{n}_1 = 200 + 0.5 \cdot 1.96^2 = 201.9208, \; \tilde{n}_2 = 100 + 0.5 \cdot 1.96^2 = 101.9208$$

Example 4

4. **Point and Interval Estimates:**
   a. Point estimate: $\hat{p}_f - \hat{p}_m = \frac{24}{200} - \frac{26}{100} = -0.14.$
   b. Confidence interval: Since $z_{0.975} = 1.96$, $y_f = 24$, $n_f = 200$, $y_m = 26$, and $n_m = 100$, the adjusted estimates of n_f and n_m are

$$\tilde{n}_1 = 200 + 0.5 \cdot 1.96^2 = 201.9208, \; \tilde{n}_2 = 100 + 0.5 \cdot 1.96^2 = 101.9208$$

The adjusted estimates of p_f and p_m are then

$$\tilde{p}_f = \frac{24 + 0.25 \cdot 1.96^2}{\tilde{n}_1} = 0.1236,$$

and

$$\tilde{p}_m = \frac{26 + 0.25 \cdot 1.96^2}{\tilde{n}_2} = 0.2645.$$

$\approx$

Example 4

The approximate score interval for $p_f - p_m$ is then

$$0.1236 - 0.2645 \pm$$

$$1.96\sqrt{\frac{0.1236(1-0.1236)}{201.9208} + \frac{0.2645(1-0.2645)}{101.9208}}$$
$$= (-0.2378, -0.0440)$$

(SAS code here)

$\approx$

Example 4

5. **Results and Interpretation:**

Example 4

5. **Results and Interpretation:** With 95% confidence we estimate that the percentage of male college students who agree or strongly agree with the statement is between 4.4 and 23.78 percent greater than the corresponding percentage of female college students.

$\approx$

Recap: Estimation: Our First Look at Statistical Inference

Recap: Estimation: Our First Look at Statistical Inference

- Population Versus Sample

Recap: Estimation: Our First Look at Statistical Inference

- Population Versus Sample
- Point Estimation

Recap: Estimation: Our First Look at Statistical Inference

- Population Versus Sample
- Point Estimation
- Sampling Distribution

Recap: Estimation: Our First Look at Statistical Inference

- Population Versus Sample
- Point Estimation
- Sampling Distribution
  - Normal

Recap: Estimation: Our First Look at Statistical Inference

- Population Versus Sample
- Point Estimation
- Sampling Distribution
  - *o* Normal
  - *o* t

Recap: Estimation: Our First Look at Statistical Inference

- Population Versus Sample
- Point Estimation
- Sampling Distribution
  - Normal
  - t
  - Binomial

Recap: Estimation: Our First Look at Statistical Inference

- Population Versus Sample
- Point Estimation
- Sampling Distribution
    - Normal
    - t
    - Binomial
- Interval Estimation

Recap: Estimation: Our First Look at Statistical Inference

- Population Versus Sample
- Point Estimation
- Sampling Distribution
  - Normal
  - t
  - Binomial
- Interval Estimation
- The Components of a Statistical Estimation Problem
  - The Scientific Goal

Recap: Estimation: Our First Look at Statistical Inference

- Population Versus Sample
- Point Estimation
- Sampling Distribution
  - Normal
  - t
  - Binomial
- Interval Estimation
- The Components of a Statistical Estimation Problem
  - The Scientific Goal
  - The Statistical Model

Recap: Estimation: Our First Look at Statistical Inference

- Population Versus Sample
- Point Estimation
- Sampling Distribution
  - *o* Normal
  - *o* t
  - *o* Binomial
- Interval Estimation
- The Components of a Statistical Estimation Problem
  - *o* The Scientific Goal
  - *o* The Statistical Model
  - *o* The Model Parameter(s) to Be Estimated

Recap: Estimation: Our First Look at Statistical Inference

- Population Versus Sample
- Point Estimation
- Sampling Distribution
  - Normal
  - t
  - Binomial
- Interval Estimation
- The Components of a Statistical Estimation Problem
  - The Scientific Goal
  - The Statistical Model
  - The Model Parameter(s) to Be Estimated
  - Point and Interval Estimates

Recap: Estimation: Our First Look at Statistical Inference

- Population Versus Sample
- Point Estimation
- Sampling Distribution
  - *o* Normal
  - *o* t
  - *o* Binomial
- Interval Estimation
- The Components of a Statistical Estimation Problem
  - *o* The Scientific Goal
  - *o* The Statistical Model
  - *o* The Model Parameter(s) to Be Estimated
  - *o* Point and Interval Estimates
  - *o* Results and Interpretation

Recap: Estimation: Our First Look at Statistical Inference

Recap: Estimation: Our First Look at Statistical Inference

- Specific Estimation Problems:

Recap: Estimation: Our First Look at Statistical Inference

- Specific Estimation Problems:
  - *o* 1-Sample Mean, Known Variance

Recap: Estimation: Our First Look at Statistical Inference

- Specific Estimation Problems:
  - *o* 1-Sample Mean, Known Variance
  - *o* 1 Sample Mean, Unknown Variance

Recap: Estimation: Our First Look at Statistical Inference

- Specific Estimation Problems:
  - *o* 1-Sample Mean, Known Variance
  - *o* 1 Sample Mean, Unknown Variance
  - *o* 1-Sample Proportion, Large Sample

Recap: Estimation: Our First Look at Statistical Inference

- Specific Estimation Problems:
  - *o* 1-Sample Mean, Known Variance
  - *o* 1 Sample Mean, Unknown Variance
  - *o* 1-Sample Proportion, Large Sample
  - *o* 1-Sample Proportion, All Sample (Approx. Score Interval)

Recap: Estimation: Our First Look at Statistical Inference

- Specific Estimation Problems:
  - *o* 1-Sample Mean, Known Variance
  - *o* 1 Sample Mean, Unknown Variance
  - *o* 1-Sample Proportion, Large Sample
  - *o* 1-Sample Proportion, All Sample (Approx. Score Interval)
  - *o* 2-Sample Mean, Paired Observations

Recap: Estimation: Our First Look at Statistical Inference

- Specific Estimation Problems:
  - *o* 1-Sample Mean, Known Variance
  - *o* 1 Sample Mean, Unknown Variance
  - *o* 1-Sample Proportion, Large Sample
  - *o* 1-Sample Proportion, All Sample (Approx. Score Interval)
  - *o* 2-Sample Mean, Paired Observations
  - *o* 2-Sample Mean, Known Variance

Recap: Estimation: Our First Look at Statistical Inference

- Specific Estimation Problems:
  - *o* 1-Sample Mean, Known Variance
  - *o* 1 Sample Mean, Unknown Variance
  - *o* 1-Sample Proportion, Large Sample
  - *o* 1-Sample Proportion, All Sample (Approx. Score Interval)
  - *o* 2-Sample Mean, Paired Observations
  - *o* 2-Sample Mean, Known Variance
  - *o* 2 Sample Mean, Unknown Variance

Recap: Estimation: Our First Look at Statistical Inference

- Specific Estimation Problems:
  - *o* 1-Sample Mean, Known Variance
  - *o* 1 Sample Mean, Unknown Variance
  - *o* 1-Sample Proportion, Large Sample
  - *o* 1-Sample Proportion, All Sample (Approx. Score Interval)
  - *o* 2-Sample Mean, Paired Observations
  - *o* 2-Sample Mean, Known Variance
  - *o* 2 Sample Mean, Unknown Variance
  - *o* 2-Sample Proportion, Large Sample

Recap: Estimation: Our First Look at Statistical Inference

- Specific Estimation Problems:
  - *o* 1-Sample Mean, Known Variance
  - *o* 1 Sample Mean, Unknown Variance
  - *o* 1-Sample Proportion, Large Sample
  - *o* 1-Sample Proportion, All Sample (Approx. Score Interval)
  - *o* 2-Sample Mean, Paired Observations
  - *o* 2-Sample Mean, Known Variance
  - *o* 2 Sample Mean, Unknown Variance
  - *o* 2-Sample Proportion, Large Sample
  - *o* 2-Sample Proportion, All Sample (Approx. Score Interval)

$\approx$

LDL Data

Subject	Baseline	Follow-up	LDL Decrease
1	160.5	168.1	-7.6
2	195.3	181.4	13.9
3	181.7	154.6	27.1
4	175.1	160.3	14.8
5	198.3	192.0	6.3
6	215.5	173.5	42.0
7	227.9	186.2	41.7
8	201.7	183.2	18.5
9	161.5	130.3	31.2
10	189.0	165.0	24.0

$\approx$