MATH1. Part II

Probability and Statistics

Chapter 9

Categorical Data: One-Sample Distribution

Variable (Review of Chapter 2)

Variable

- **Variable:** a variable is a <u>characteristic</u> of a person or a thing that can be assigned a number or a category.
  - For example, blood type (A, B, AB, O) and age are two variables we might measure on a person.
- Types of variables:
  - A **categorical variable** is a variable that records which of several categories a person or thing is in.
  - A **numeric variable** records the amount of something.
    - A **continuous variable** is a numeric variable that is measured on a continuous scale.
    - Some types of numeric variables are not continuous but fall on a discrete scale, with spaces between the possible values. A **discrete variable** is a numeric variable for which we can list the possible values.

What are the type of variables we were studying?

Chapter 4 The Normal Distribution

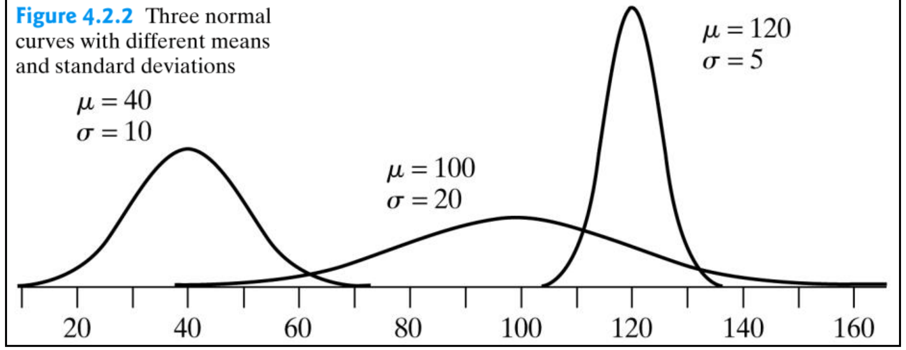


Figure 4.2.2 Three normal curves with different means and standard deviations

Chapter 5 Sampling Distributions

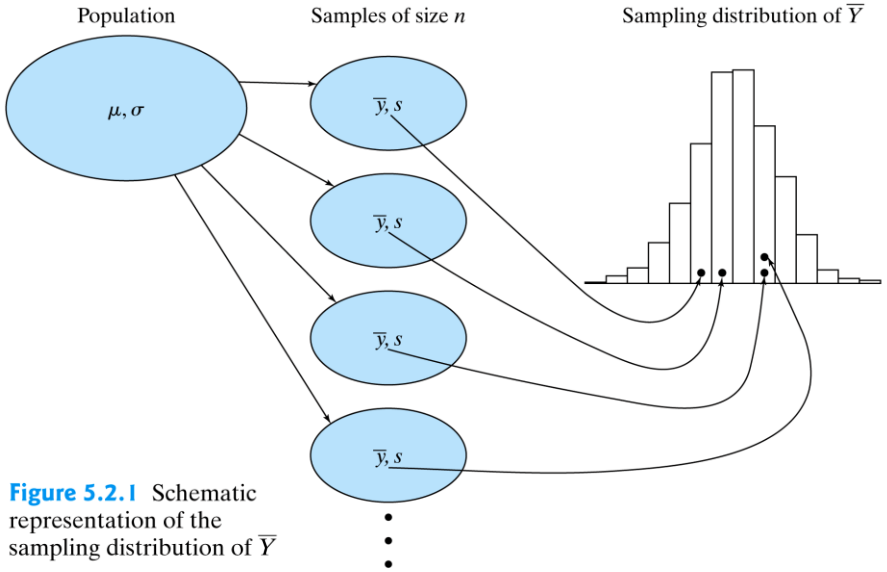


Figure 5.2.1 Schematic representation of the sampling distribution of $\bar{Y}$

Chapter 6 Confidence Intervals

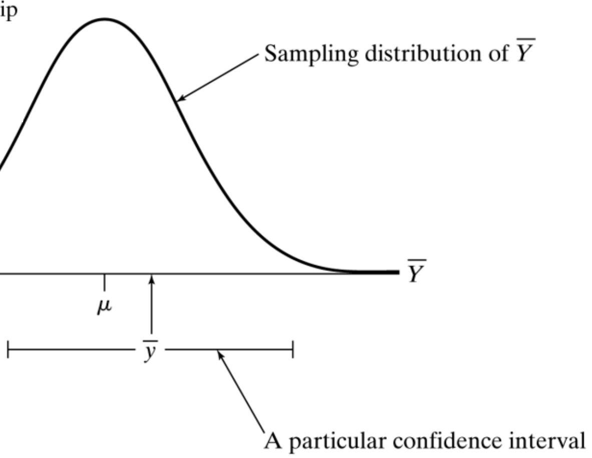


Figure 6.3.7 Relationship between a particular confidence interval for μ and the sampling distribution of $\bar{Y}$

Chapter 7/8

Comparison of Two Independent / Paired Samples

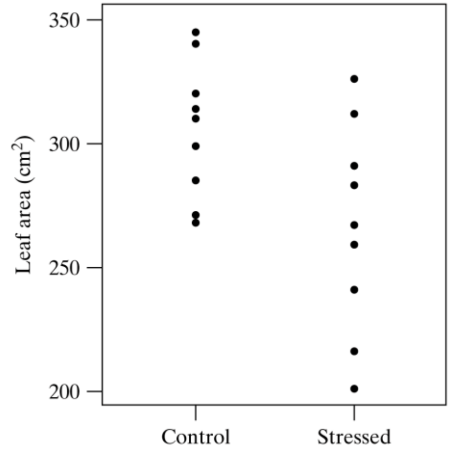


Figure 7.1.2 Parallel dotplots of leaf areas

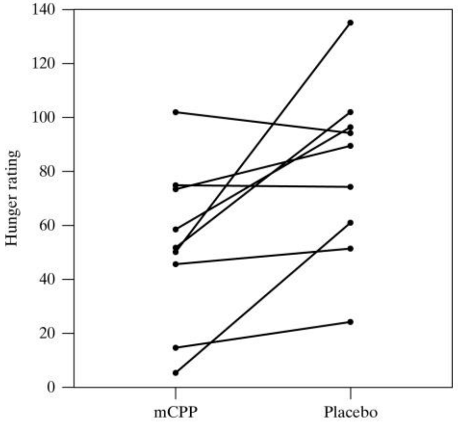


Figure 8.1.1 Dotplots of hunger ratings after mCPP and placebo, with line segments connecting readings on each subject

What are the type of variables we were studying?

Chapter 4 The Normal Distribution

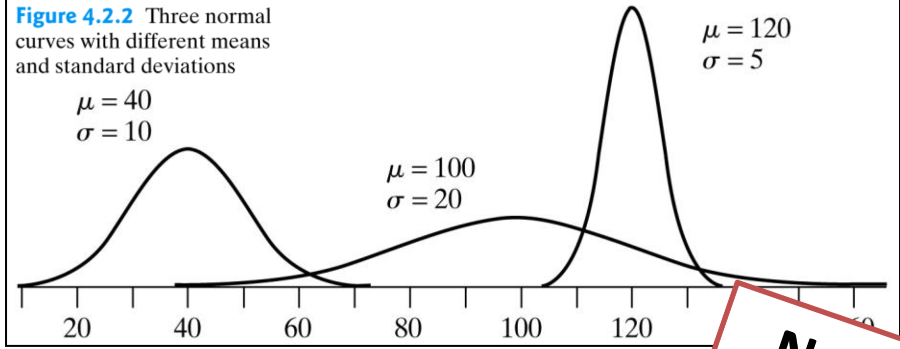


Figure 4.2.2 Three normal curves with different means and standard deviations

Chapter 5 Sampling Distributions

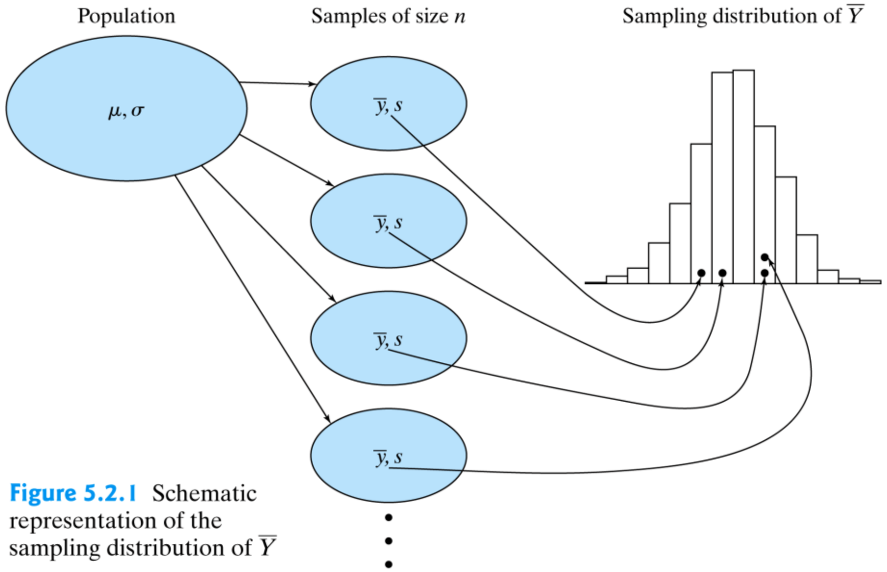


Figure 5.2.1 Schematic representation of the sampling distribution of $\bar{Y}$

Chapter 6 Confidence Intervals

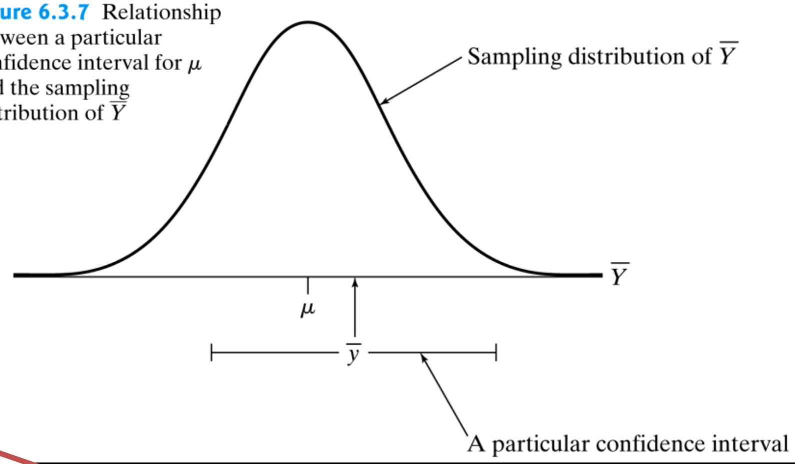


Figure 6.3.7 Relationship between a particular confidence interval for μ and the sampling distribution of $\bar{Y}$

Numeric Variable

Chapter 7/8

Comparison of Two Independent / Paired Samples

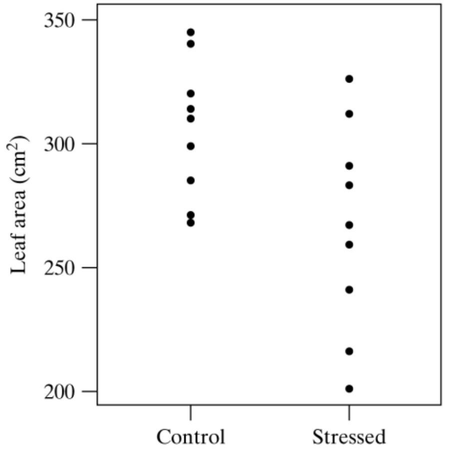


Figure 7.1.2 Parallel dotplots of leaf areas

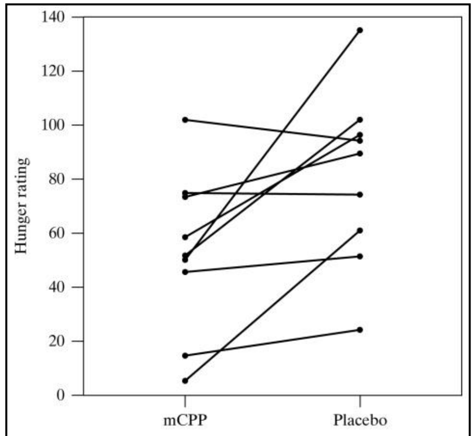


Figure 8.1.1 Dotplots of hunger ratings after mCPP and placebo, with line segments connecting readings on each subject

9.1 Dichotomous Observations

Dichotomous categorical variable

- **Dichotomous categorical variable** is defined as a categorical variable that has only two possible values.
- Examples of dichotomous variables
  - Heads or Tails.
  - Male or Female.
  - Under age 65 or 65 and over.

Example 9.1.1 Contaminated Soda

- To estimate the proportion of contaminated soft-drink dispensers in a community in Virginia, researchers randomly sampled 30 dispensers and found 5 to be contaminated with *Chryseobacterium meningosepticum*.
- What is the dichotomous variables in this question?

9.1 Dichotomous Observations

Dichotomous categorical variable

- **Dichotomous categorical variable** is defined as a categorical variable that has only two possible values.
- Examples of dichotomous variables
  - Heads or Tails.
  - Male or Female.
  - Under age 65 or 65 and over.

Example 9.1.1 Contaminated Soda

- To estimate the proportion of contaminated soft-drink dispensers in a community in Virginia, researchers randomly sampled 30 dispensers and found 5 to be contaminated with *Chryseobacterium meningosepticum*.
- What is the dichotomous variables in this question?
  - dichotomous variables: contaminated or not contaminated

9.1 Dichotomous Observations

Population proportion, *p*

- For a categorical variable, we can describe a population by simply stating the proportion, or relative frequency, of the population in each category.
  - For categorical data, the sample proportion $\hat{p}$ (p-hat) of a category is an estimate of the corresponding population proportion.
  - The Wilson-adjusted sample proportion, $\tilde{p}$ (p-tilde), is another estimate of the population proportion.

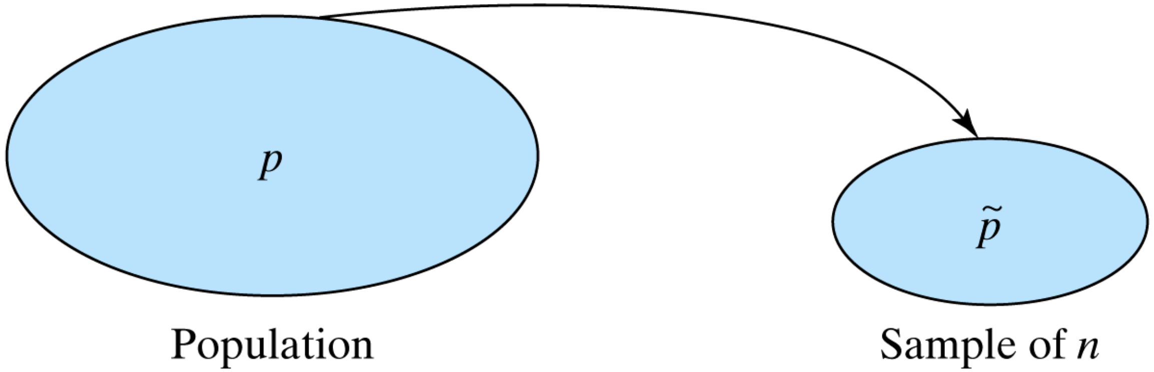


Figure 9.2.1 Notation for population and sample proportion

9.1 Dichotomous Observations

Sample Proportion, $\hat{p}$

- For categorical data, **Sample proportion**, $\hat{p}$ (p-hat), is defined as:

$$\hat{p} = y/n,$$

  - where y is the number of observations in the sample with the attribute of interest, and
  - n is the sample size.

Example 9.1.1 Contaminated Soda (continued)

- To estimate the proportion of contaminated soft-drink dispensers in a community in Virginia, researchers randomly sampled 30 dispensers and found 5 to be contaminated with *Chryseobacterium meningosepticum*.
- What is the sample proportion?

9.1 Dichotomous Observations

Sample Proportion, $\hat{p}$

- For categorical data, **Sample proportion**, $\hat{p}$ (p-hat), is defined as:

$$\hat{p} = y/n,$$

  - where y is the number of observations in the sample with the attribute of interest, and
  - n is the sample size.

Example 9.1.1 Contaminated Soda (continued)

- To estimate the proportion of contaminated soft-drink dispensers in a community in Virginia, researchers randomly sampled 30 dispensers and found 5 to be contaminated with *Chryseobacterium meningosepticum*.
- What is the sample proportion?
  - Sample proportion, $\hat{p} = y/n = 5/30 = 0.167$

9.1 Dichotomous Observations

Wilson-adjusted sample proportion, $\tilde{p}$

- For categorical data, **Wilson-adjusted sample proportion**, $\tilde{p}$ (p-tilde), is

$$\tilde{p} = (y+2) / (n+4)$$

  - where y is the number of observations in the sample with the attribute of interest, and
  - n is the sample size.
  - This augmentation has the effect of biasing the estimate towards the value 1/2.
  - https://stats.stackexchange.com/questions/109429/wilsons-adjustment-for-sample-proportion

Example 9.1.1 Contaminated Soda (continued)

- To estimate the proportion of contaminated soft-drink dispensers in a community in Virginia, researchers randomly sampled 30 dispensers and found 5 to be contaminated with *Chryseobacterium meningosepticum*.
- What is the Wilson-adjusted sample proportion?

9.1 Dichotomous Observations

Wilson-adjusted sample proportion, $\tilde{p}$

- For categorical data, **Wilson-adjusted sample proportion**, $\tilde{p}$ (p-tilde), is

$$\tilde{p} = (y+2) / (n+4)$$

  – where y is the number of observations in the sample with the attribute of interest, and
  – n is the sample size.
  – This augmentation has the effect of biasing the estimate towards the value 1/2.

Example 9.1.1 Contaminated Soda (continued)

- To estimate the proportion of contaminated soft-drink dispensers in a community in Virginia, researchers randomly sampled 30 dispensers and found 5 to be contaminated with *Chryseobacterium meningosepticum*.
- What is the Wilson-adjusted sample proportion?
  – Wilson-adjusted sample proportion, $\tilde{p} = (y+2) / (n+4) = (5 + 2) / (30 + 4) = 0.206$

9.1 Dichotomous Observations

Wilson-adjusted sample proportion, $\tilde{p}$

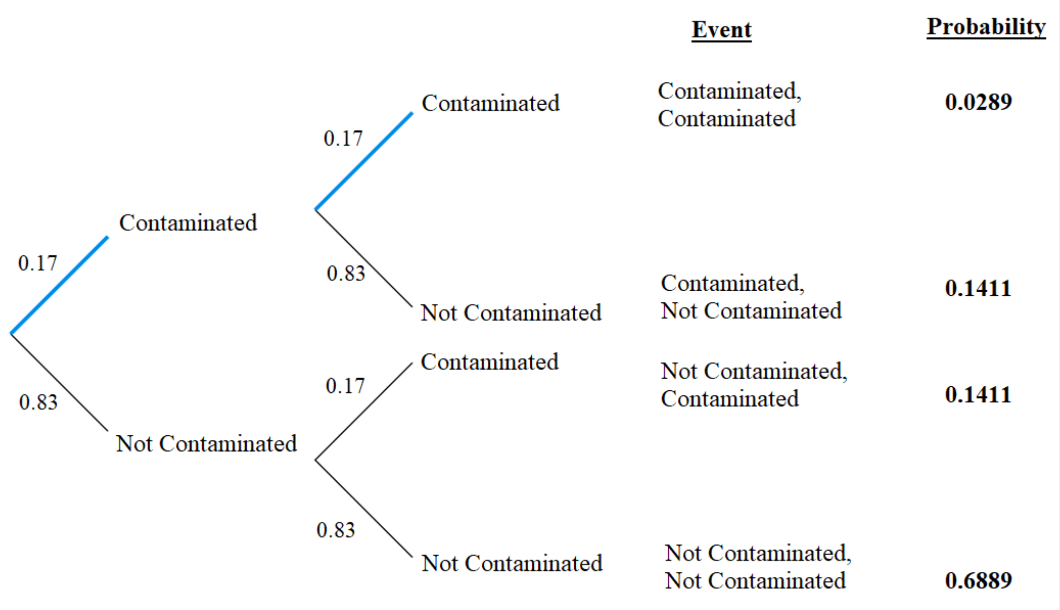
- For categorical data, **Wilson-adjusted sample proportion**, $\tilde{p}$ (p-tilde), is

$$\tilde{p} = (y+2) / (n+4)$$

Example 9.1.3 Contaminated Soda (continued)

- 17% of all soft-drink dispensers are contaminated
- examine a random sample of **2** drink dispensers from this population of dispensers

- What is the Wilson-adjusted sample proportion of contaminated dispensers?



9.1 Dichotomous Observations

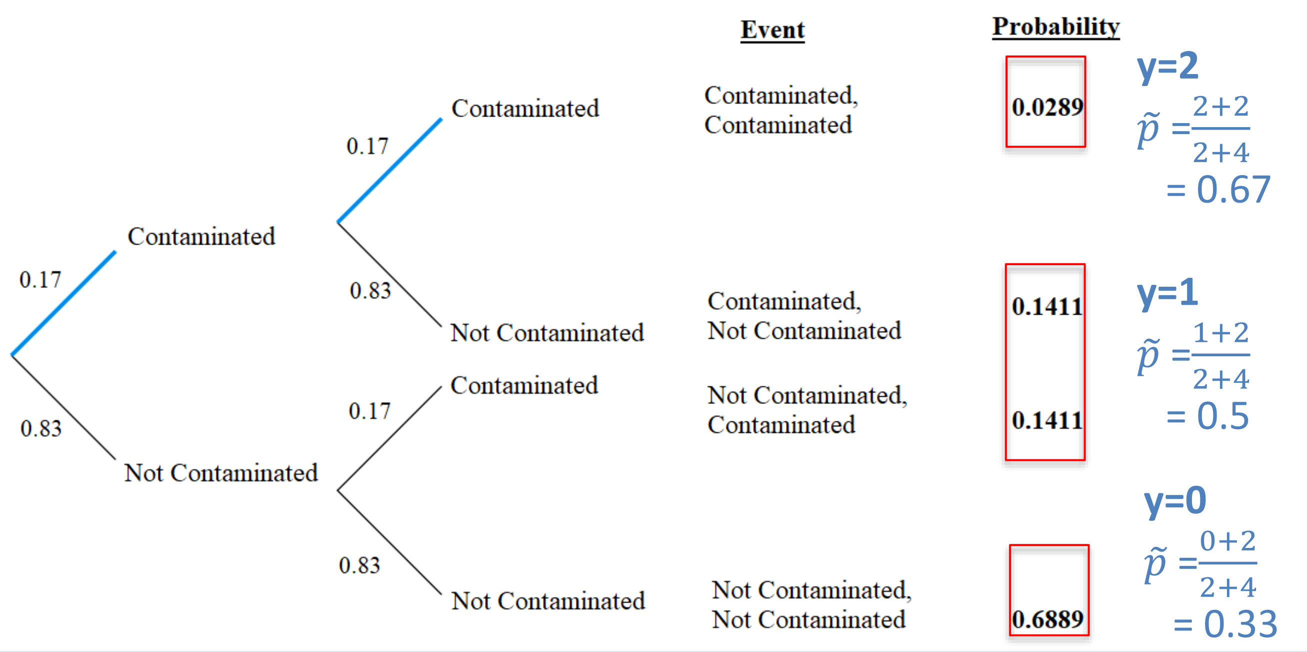
Wilson-adjusted sample proportion, $\tilde{p}$

- For categorical data, **Wilson-adjusted sample proportion**, $\tilde{p}$ (p-tilde), is

$$\tilde{p} = (y+2) / (n+4)$$

Example 9.1.3 Contaminated Soda (continued)

- 17% of all soft-drink dispensers are contaminated
- examine a random sample of **2** drink dispensers from this population of dispensers
- What is the Wilson-adjusted sample proportion of contaminated dispensers?
  - $\tilde{p} = (y+2) / (n+4)$



9.1 Dichotomous Observations

The Sampling Distribution of $\tilde{P}$

Example 9.1.3 Contaminated Soda (continued)

- 17% of all soft-drink dispensers are contaminated
- examine a random sample of **2** drink dispensers from this population of dispensers

- **Sampling Distribution of Y and $\tilde{P}$**

Table 9.1.1 Sampling distribution of Y (the number of contaminated dispensers) and of $\tilde{P}$ (the Wilson-adjusted proportion of contaminated dispensers) for samples of size $n = 2$ for a population with 17% of the dispensers contaminated

Y	$\tilde{P}$	Probability
0	0.33	0.6889
1	0.50	0.2822
2	0.67	0.0289

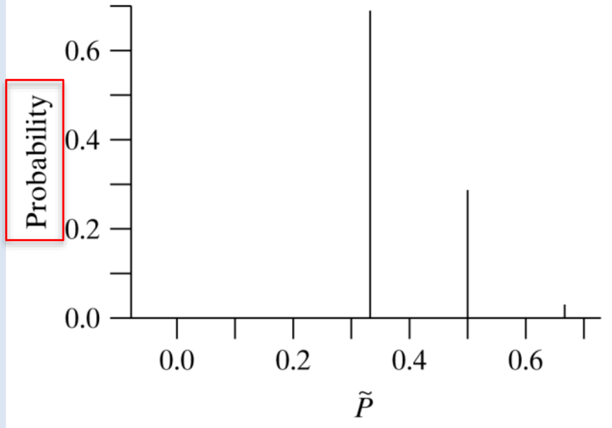


Figure 9.1.1 Sampling distribution of $\tilde{P}$ for $n = 2$ and $p = 0.17$

Probability

0.0289 — y=2, $\tilde{p}$ = 0.67

0.1411, 0.1411 — y=1, $\tilde{p}$ = 0.5

0.6889 — y=0, $\tilde{p}$ = 0.33

9.1 Dichotomous Observations

The Sampling Distribution of $\tilde{P}$

Example 9.1.3 Contaminated Soda (continued)

- 17% of all soft-drink dispensers are contaminated
- examine a random sample of **20** drink dispensers from this population of dispensers
- What is the Sampling Distribution of Y and $\tilde{P}$?

9.1 Dichotomous Observations

The Sampling Distribution of $\tilde{P}$

Example 9.1.3 Contaminated Soda (continued)

- 17% of all soft-drink dispensers are contaminated
- examine a random sample of **20** drink dispensers from this population of dispensers
- What is the Sampling Distribution of Y and $\tilde{P}$?
  - Wilson-adjusted sample proportion
    $\tilde{p}$ = (y+2) / (n+4) = (0 + 2) / (20 + 4) = 0.0833
  - Binomial Distribution (Review of Chapter 3)
    $\Pr\{j\ successes\} = \Pr(Y = j\} = {}_nC_j p^j (1-p)^{n-j}$ = Pr { Y = 0) = 0.0241

9.1 Dichotomous Observations

The Sampling Distribution of $\tilde{P}$

Example 9.1.3 Contaminated Soda (continued)

- 17% of all soft-drink dispensers are contaminated
- examine a random sample of **20** drink dispensers from this population of dispensers
- What is the Sampling Distribution of Y and $\tilde{P}$?

Table 9.1.2 Sampling distribution of Y, the number of successes, and of $\tilde{P}$, the Wilson-adjusted proportion of successes, when $n = 20$ and $p = 0.17$

Y	$\tilde{P}$	Probability	Y	$\tilde{P}$	Probability
0	0.0833	0.0241	11	0.5417	0.0001
1	0.1250	0.0986	12	0.5833	0.0000
2	0.1667	0.1919	13	0.6250	0.0000
3	0.2083	0.2358	14	0.6667	0.0000
4	0.2500	0.2053	15	0.7083	0.0000
5	0.2917	0.1345	16	0.7500	0.0000
6	0.3333	0.0689	17	0.7917	0.0000
7	0.3750	0.0282	18	0.8333	0.0000
8	0.4167	0.0094	19	0.8750	0.0000
9	0.4583	0.0026	20	0.9167	0.0000
10	0.5000	0.0006			

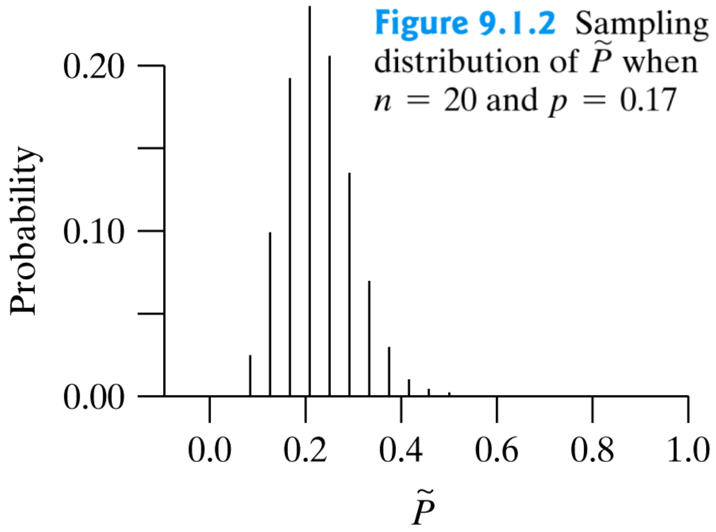
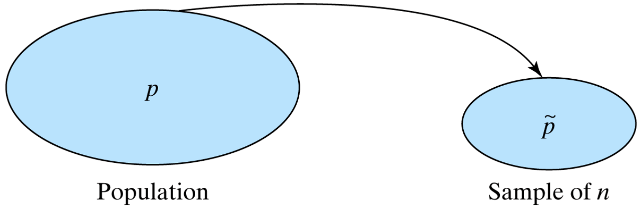


Figure 9.1.2 Sampling distribution of $\tilde{P}$ when $n = 20$ and $p = 0.17$

9.1 Dichotomous Observations

Relationship to statistical inference

- $\tilde{P}$ as our estimate of p.
- The sampling distribution of $\tilde{P}$ can be used to predict how much sampling error to expect in this estimate.



Example 9.1.3 Contaminated Soda (continued)

- 17% of all soft-drink dispensers are contaminated
- examine a random sample of **20** drink dispensers from this population of dispensers
- What is the probability that $\tilde{P}$ will be within ± 0.05 of p?

Table 9.1.2 Sampling distribution of Y, the number of successes, and of $\tilde{P}$, the Wilson-adjusted proportion of successes, when $n = 20$ and $p = 0.17$

Y	$\tilde{P}$	Probability	Y	$\tilde{P}$	Probability
0	0.0833	0.0241	11	0.5417	0.0001
1	0.1250	0.0986	12	0.5833	0.0000
2	0.1667	0.1919	13	0.6250	0.0000
3	0.2083	0.2358	14	0.6667	0.0000
4	0.2500	0.2053	15	0.7083	0.0000
5	0.2917	0.1345	16	0.7500	0.0000
6	0.3333	0.0689	17	0.7917	0.0000
7	0.3750	0.0282	18	0.8333	0.0000
8	0.4167	0.0094	19	0.8750	0.0000
9	0.4583	0.0026	20	0.9167	0.0000
10	0.5000	0.0006			

9.1 Dichotomous Observations

The Sampling Distribution of $\tilde{P}$

Example 9.1.3 Contaminated Soda (continued)

- 17% of all soft-drink dispensers are contaminated
- examine a random sample of **20** drink dispensers from this population of dispensers
- What is the probability that $\tilde{P}$ will be within $\pm$0.05 of p?
  - $\Pr\{(0.17-0.05) \le \tilde{P} \le (0.17+0.05)\}$

    $= \Pr\{0.12 \le \tilde{P} \le 0.22\}$

    $= 0.0986 + 0.1919 + 0.2358$

    $= 0.5263$

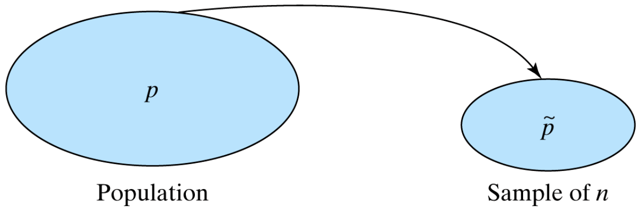
Table 9.1.2 Sampling distribution of Y, the number of successes, and of $\tilde{P}$, the Wilson-adjusted proportion of successes, when $n = 20$ and $p = 0.17$

Y	$\tilde{P}$	Probability	Y	$\tilde{P}$	Probability
0	0.0833	0.0241	11	0.5417	0.0001
1	0.1250	0.0986	12	0.5833	0.0000
2	0.1667	0.1919	13	0.6250	0.0000
3	0.2083	0.2358	14	0.6667	0.0000
4	0.2500	0.2053	15	0.7083	0.0000
5	0.2917	0.1345	16	0.7500	0.0000
6	0.3333	0.0689	17	0.7917	0.0000
7	0.3750	0.0282	18	0.8333	0.0000
8	0.4167	0.0094	19	0.8750	0.0000
9	0.4583	0.0026	20	0.9167	0.0000
10	0.5000	0.0006			

9.1 Dichotomous Observations

Dependence on sample size

- The larger the value of n, then the more likely it is $\tilde{P}$ will be close to p.



Example 9.1.3 Contaminated Soda

- 17% of all soft-drink dispensers are contaminated
- examine a random sample of **n** drink dispensers from this population of dispensers
- What is the Sampling Distribution of $\tilde{P}$?
  - — Figures 9.1.3

Table 9.1.3

n	$\Pr\{0.12 \le \tilde{P} \le 0.22\}$
20	0.53
40	0.56
80	0.75
400	0.99

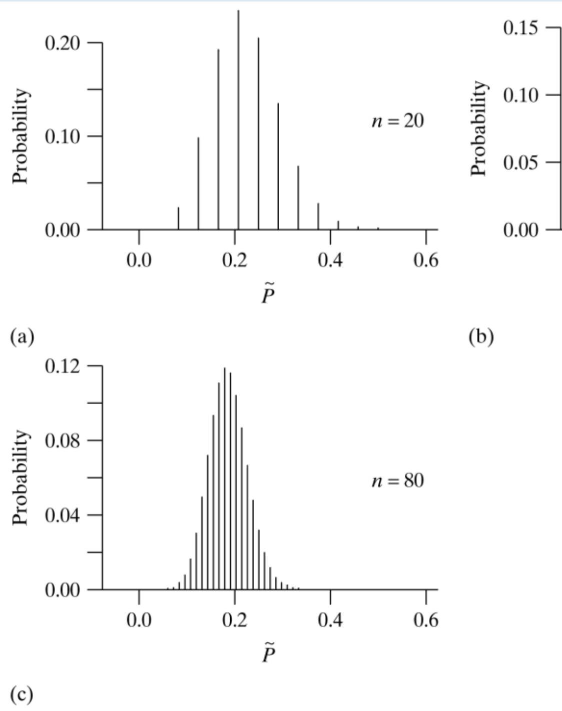


Figure 9.1.3 Sampling distributions of $\tilde{P}$ for $p = 0.17$ and various values of n

To be continued ….

Tomorrow

9.2 Confidence Interval for a Population Proportion

Confidence Interval for μ (Review of Chapter 6)

STANDARD ERROR OF THE MEAN

$$\mathrm{SE}_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

CONFIDENCE INTERVAL FOR μ

95% confidence interval: $\bar{y} \pm t_{0.025}\mathrm{SE}_{\bar{Y}}$

Critical value $t_{0.025}$ from Student's t distribution with df $= n - 1$.

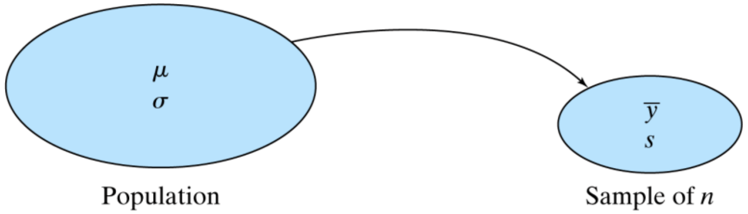


Figure 6.1.1 Notation for means and SDs of sample and population

Confidence interval for p

Standard Error of $\tilde{p}$ (for a 95% Confidence Interval)

$$\mathrm{SE}_{\tilde{p}} = \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{n + 4}}$$

95% Confidence Interval for p

95% confidence interval: $\tilde{p} \pm 1.96\mathrm{SE}_{\tilde{p}}$, where $\tilde{p}$ = (y+2) / (n+4)

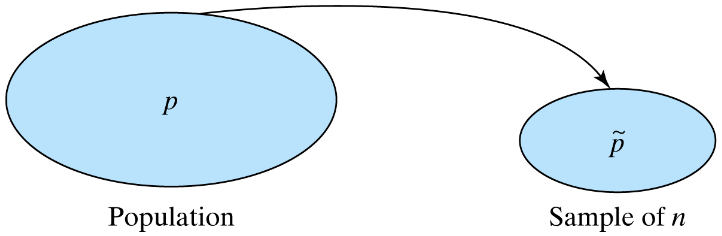


Figure 9.2.1 Notation for population and sample proportion

9.2 Confidence Interval for a Population Proportion

Standard error of $\tilde{P}$

$$SE_{\tilde{p}} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}, \text{ where } \tilde{p} = (y+2) / (n+4)$$

95% Confidence interval for p

$$\tilde{p} \pm t_{0.025} SE_{\tilde{p}} \quad \rightarrow \quad \tilde{p} \pm 1.96\, SE_{\tilde{p}}$$

Example 9.2.2 breast cancer

- BRCA1 is a gene that has been linked to breast cancer.
- Of the 169 women tested, 27 (16%) had BRCA1 mutations.
- Let p denote the probability that a woman with a family history of breast cancer will have a BRCA1 mutation.
- What is the 95% confidence interval for p?

9.2 Confidence Interval for a Population Proportion

Confidence interval for p

Example 9.2.2 breast cancer

- BRCA1 is a gene that has been linked to breast cancer.
- Of the 169 women tested, 27 (16%) had BRCA1 mutations.
- Let p denote the probability that a woman with a family history of breast cancer will have a BRCA1 mutation.
- What is the 95% confidence interval for p?
  - $\tilde{p}$ = (27 + 2) / (169 + 4) = 0.168
  - Standard error of $\tilde{P}$: $SE_{\tilde{p}} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}} = \sqrt{\frac{0.168\,(1-0.168)}{169+4}} = 0.028$
  - 95% Confidence interval for p:

    $\tilde{p} \pm t_{0.025}\, SE_{\tilde{p}} \rightarrow 0.168 \pm 1.96 \times 0.028 \rightarrow 0.113 < p < 0.223$
  - Thus, we are 95% confident that the probability of a BRCA1 mutation in a woman with a family history of breast cancer is between 11.3% and 22.3%.

9.2 Confidence Interval for a Population Proportion

Conditions for use of the Wilson 95% confidence interval for p

$$\tilde{p} \pm t_{0.025} SE_{\tilde{p}} \quad \rightarrow \quad \tilde{p} \pm 1.96\, SE_{\tilde{p}}$$

- Regard the data as a <u>random sample</u> from some population.
- The Wilson interval does not require large sample sizes to be valid.

One-sided confidence interval

- one-sided 95% (upper) confidence interval Pr ($-\infty < p < \tilde{p} + t_{0.05} SE_{\tilde{p}}$)
  - we are 95% confident that the probability of p is at most $\tilde{p} + t_{0.05} SE_{\tilde{p}}$
- one-sided 95% (lower) confidence interval Pr ($\tilde{p} - t_{0.05} SE_{\tilde{p}} < p < +\infty$)
  - we are 95% confident that the probability of p is at least $\tilde{p} - t_{0.05} SE_{\tilde{p}}$

9.2 Confidence Interval for a Population Proportion

Planning a study to estimate p

$$\text{Desired SE} = \sqrt{\frac{(Guessed\ \tilde{p})(1-Guessed\ \tilde{p})}{n+4}}$$

– Where Guessed $\tilde{p}$ = (y+2) / (n+4)

Example 9.2.6 Vegetarians

- In a survey of 136 students at a U.S. college, 19 of them said that they were vegetarians.
- The sample estimate of the proportion is

$$\tilde{p} = (19 + 2) / (136 + 4) = 0.15$$

- Suppose we regard these data as a pilot study and we now wish to plan a study large enough to estimate *p* with a standard error of two percentage points, that is, **0.02**.
- What is the minimal sample size n?

9.2 Confidence Interval for a Population Proportion

Planning a study to estimate p

Example 9.2.6 Vegetarians

- In a survey of 136 students at a U.S. college, 19 of them said that they were vegetarians.
- The sample estimate of the proportion is

$$\tilde{p} = (19 + 2) / (136 + 4) = 0.15$$

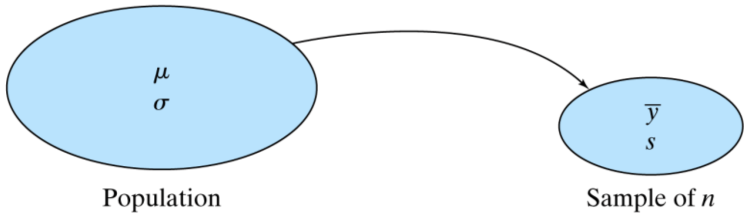
- Suppose we regard these data as a pilot study and we now wish to plan a study large enough to estimate *p* with a standard error of two percentage points, that is, 0.02.
- What is the minimal sample size n?
  - We choose *n* to satisfy the following relation:
  - Desired SE = $\sqrt{\frac{(Guessed\ \tilde{p})(1-Guessed\ \tilde{p})}{n+4}} = \sqrt{\frac{0.15(0.85)}{n+4}} \leq 0.02$
  - This equation is easily solved to give $n + 4 \geq 318.75$.
  - We should plan a sample of 315 students.

9.4 Inference for Proportions

Hypothesis Testing: The t Test (Review of Chapter 7)

- Null hypothesis: H_0: μ1 = μ2
- Alternative hypothesis: H_A: μ1 ≠ μ2
- The t test test statistic: $t_s = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{SE_{(\bar{Y}_1 - \bar{Y}_2)}}$ => P-value vs. α

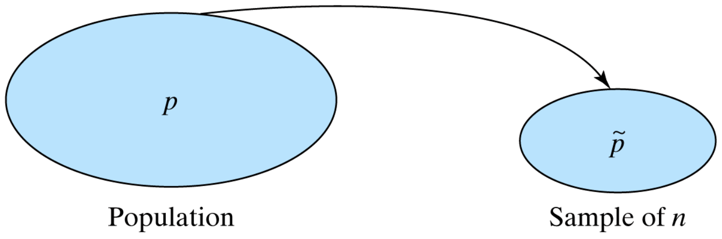
Figure 6.1.1 Notation for means and SDs of sample and population



Hypothesis Testing: The Chi-Square Goodness-of-Fit Test

- Null hypothesis: H_0
- Alternative hypothesis: H_A
- The Chi-square statistic: χ^2_s => P-value vs. α

Figure 9.2.1 Notation for population and sample proportion

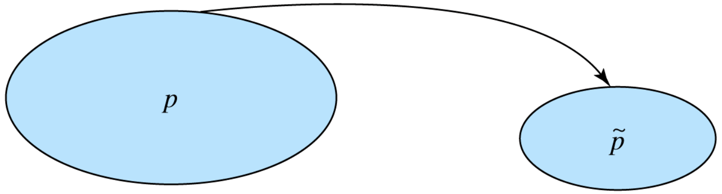


9.4 Inference for Proportions

The Chi-Square Goodness-of-Fit Test

- Hypothesis testing for <u>categorical data</u>
- $\tilde{P}$ as our estimate of p.
- The sampling distribution of $\tilde{P}$ can be used to predict how much sampling error to expect in this estimate.

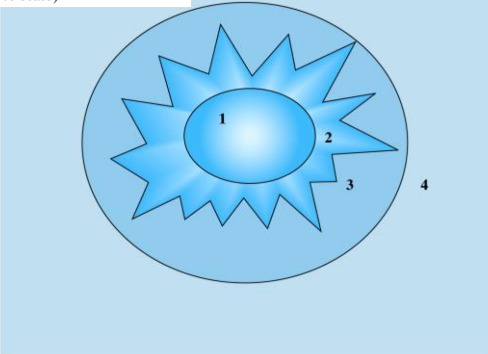
Figure 9.2.1 Notation for population and sample proportion



Example 9.4.1 Deer Habitat and Fire

- Overall: 3,000 acres, 75 deer
- Does fire affect deer behavior?

Figure 9.4.1 Schematic of 3,000-acre parcel with an interior 730-acre fire (not to scale)



1. the region near the heat of the burn, 520 acres, 2 deer
2. the inside edge of the burn, 210 acres, 12 deer
3. the outside edge of the burn, 240 acres, 18 deer
4. the area outside of the burned area, 2030 acres, 43 deer

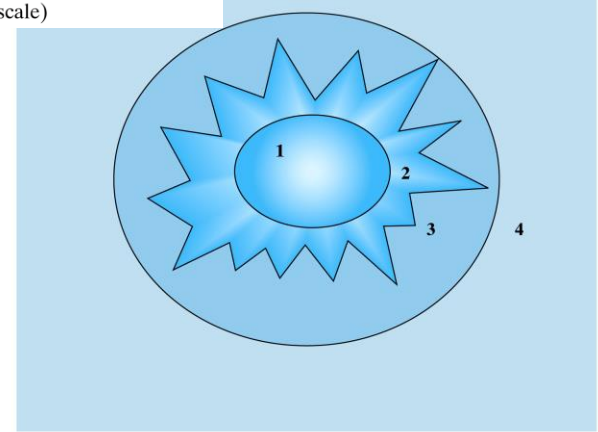
9.4 Inference for Proportions

The Chi-Square Goodness-of-Fit Test

Example 9.4.1 Deer Habitat and Fire

- Overall: 3,000 acres, 75 deer
- Does fire affect deer behavior?
  - — H_0: deer show no preference to any particular type of burned/unburned habitat (deer are randomly distributed over the 3,000 acres).
  - — H_A: deer do show a preference for some of the regions (deer are NOT randomly distributed over the 3,000 acres).

Figure 9.4.1 Schematic of 3,000-acre parcel with an interior 730-acre fire (not to scale)



1. the region near the heat of the burn, 520 acres, 2 deer
2. the inside edge of the burn, 210 acres, 12 deer
3. the outside edge of the burn, 240 acres, 18 deer
4. the area outside of the burned area, 2030 acres, 43 deer

9.4 Inference for Proportions

The Chi-Square Goodness-of-Fit Test

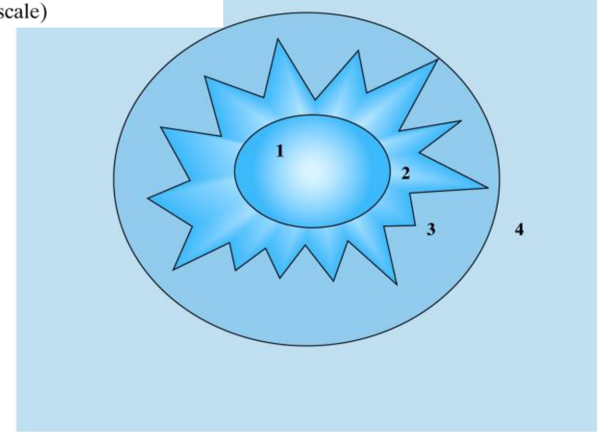
Example 9.4.1 Deer Habitat and Fire

- Overall: 3,000 acres, 75 deer
- Does fire affect deer behavior?
  - H_0: deer are randomly distributed over the 3,000 acres.
    - Pr {inner burn} = 520/3000 = 0.173
    - Pr {inner edge} = 210/3000 = 0.070
    - Pr {outer edge} = 240/3000 = 0.080
    - Pr {outer unburned} =2030/3000 = 0.677

Compound null hypothesis:

- a goodness-of-fit null hypothesis can contain more than one assertion.
- Such a null hypothesis called a compound null hypothesis.
  - 1. alternative hypothesis is necessarily nondirectional
  - 2. if H_0 is rejected, the test does not yield a directional conclusion

Figure 9.4.1 Schematic of 3,000-acre parcel with an interior 730-acre fire (not to scale)



1. the region near the heat of the burn, 520 acres, 2 deer
2. the inside edge of the burn, 210 acres, 12 deer
3. the outside edge of the burn, 240 acres, 18 deer
4. the area outside of the burned area, 2030 acres, 43 deer

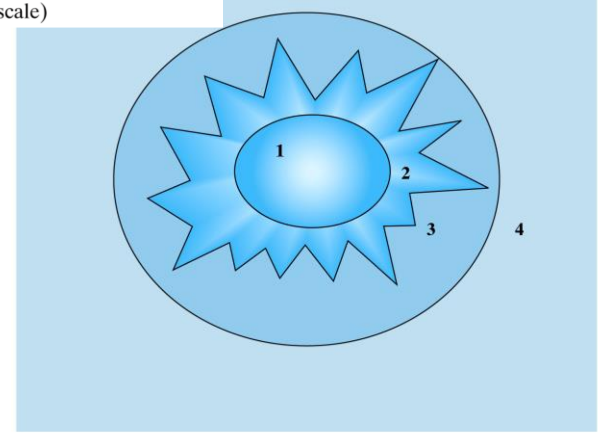
9.4 Inference for Proportions

The Chi-Square Goodness-of-Fit Test

Example 9.4.1 Deer Habitat and Fire

- Overall: 3,000 acres, 75 deer
- Does fire affect deer behavior?
  - **H_0**: deer are randomly distributed over the 3,000 acres.
    - Pr {inner burn} = 520/3000 = 0.173; e_1 = 0.173x75 = 13.00
    - Pr {inner edge} = 210/3000 = 0.070; e_2 = 0.070x75 = 5.25
    - Pr {outer edge} = 240/3000 = 0.080; e_3 = 0.080x75 = 6.00
    - Pr {outer unburned} =2030/3000 = 0.677; e_4 = 0.677x75 = 50.75
  - **H_A**: deer are NOT randomly distributed over the 3,000 acres.
    - Pr {inner burn} ≠ 0.173
    - Pr {inner edge} ≠ 0.070
    - Pr {outer edge} ≠ 0.080
    - Pr {outer unburned} ≠ 0.677

Figure 9.4.1 Schematic of 3,000-acre parcel with an interior 730-acre fire (not to scale)



1. the region near the heat of the burn, 520 acres, 2 deer
2. the inside edge of the burn, 210 acres, 12 deer
3. the outside edge of the burn, 240 acres, 18 deer
4. the area outside of the burned area, 2030 acres, 43 deer

9.4 Inference for Proportions

The Chi-Square Goodness-of-Fit Test

- **Chi-square statistic**:

$$\chi^2_s = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

— o_i - observed frequency of category i,
— e_i - expected frequency of category i,
— where the summation is over all k categories.

— small values of χ^2_s would indicate that the data agree with H_0, while large values of χ^2_s would indicate disagreement.

— if the sample size is large enough, then the null distribution of χ^2_s can be approximated by a distribution known as a χ^2 distribution (**Table 9**).

— df = k - 1, where k equals the number of categories.

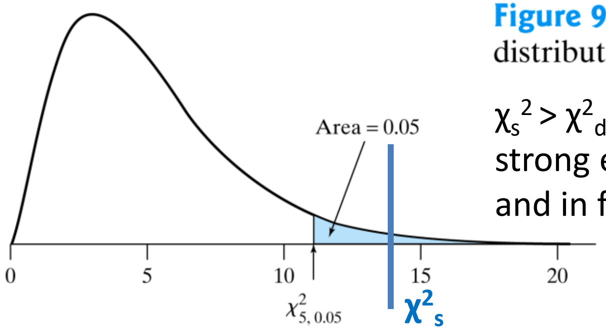


Figure 9.4.4 The χ^2 distribution with df = 5

$\chi_s^2 > \chi^2_{df,\,0.05}$, P-value < 0.05, strong evidence against H_0 and in favor of H_A

TABLE 9 Critical Values of the Chi-Square Distribution

Note: Column headings are non-directional (omni-directional) *P*-values. If H_A is directional (which is only possible when df = 1), the directional *P*-values are found by dividing the column headings in half.

	TAIL PROBABILITY						
df	0.20	0.10	0.05	0.02	0.01	0.001	0.0001
1	1.64	2.71	3.84	5.41	6.63	10.83	15.14
2	3.22	4.61	5.99	7.82	9.21	13.82	18.42
3	4.64	6.25	7.81	9.84	11.34	16.27	21.11

9.4 Inference for Proportions

The Chi-Square Goodness-of-Fit Test

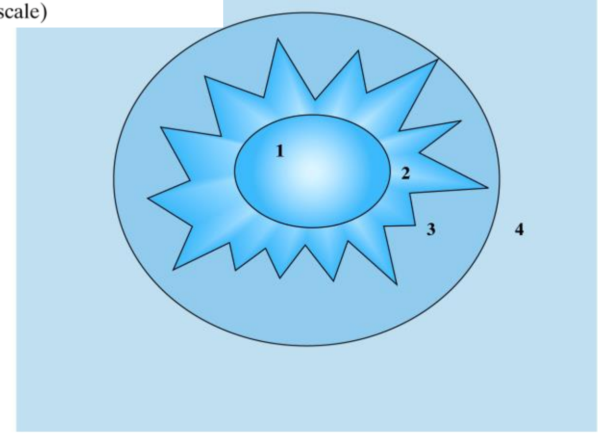
Example 9.4.1 Deer Habitat and Fire

- Overall: 3,000 acres, 75 deer
- Does fire affect deer behavior?
  - **H_0**: deer are randomly distributed over the 3,000 acres.
    - Pr {inner burn} = 520/3000 = 0.173; e_1 = 0.173x75 = 13.00
    - Pr {inner edge} = 210/3000 = 0.070; e_2 = 0.070x75 = 5.25
    - Pr {outer edge} = 240/3000 = 0.080; e_3 = 0.080x75 = 6.00
    - Pr {outer unburned} =2030/3000 = 0.677; e_4 = 0.677x75 = 50.75
  - $\chi^2_s = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$

    $= \frac{(2-13)^2}{13} + \frac{(12-5.25)^2}{5.25} + \frac{(18-6)^2}{6} + \frac{(43-50.75)^2}{50.75}$

    $= 43.2$
  - $\chi^2_s > \chi^2_{3,\ 0.0001}$ (Table 9); P-value < 0.0001 < 0.05
  - We have strong evidence against H_0 and in favor of H_A

Figure 9.4.1 Schematic of 3,000-acre parcel with an interior 730-acre fire (not to scale)



1. the region near the heat of the burn, 520 acres, 2 deer (o_1)
2. the inside edge of the burn, 210 acres, 12 deer (o_2)
3. the outside edge of the burn, 240 acres, 18 deer (o_3)
4. the area outside of the burned area, 2030 acres, 43 deer (o_4)

9.4 Inference for Proportions

The Chi-Square Test (χ^2 test)

Goodness-of-fit test

Data:

$$o_i = \text{the observed frequency of category } i$$

Null hypothesis:

H_0 specifies the probability of each category.*

Calculation of expected frequencies:

$$e_i = n \times \text{Probability specified for category } i \text{ by } H_0$$

Test statistic:

$$\chi_s^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

Null distribution (approximate):

$$\chi^2 \text{ distribution with df} = k - 1$$

where k = the number of categories

This approximation is adequate if $e_i \geq 5$ for every category.

9.4 Inference for Proportions

Directional Alternative

- A chi-square goodness-of-fit test against a directional alternative (when the observed variable is dichotomous) uses the familiar two-step procedure:
- **Step 1**. Check directionality (see if the data deviate from H_0 in the direction specified by H_A).
  - (a) If not, the P-value is greater than 0.50.
  - (b) If so, proceed to step 2.
- **Step 2**. The P-value is half what it would be if H_A were nondirectional.

Example 9.4.9 Harvest Moon Festival

- Can people who are close to death postpone dying until after a symbolically meaningful occasion?

Table 9.4.4 Harvest moon festival data

	Before	After	Total
Observed	33	70	103
Expected	51.5	51.5	103

9.4 Inference for Proportions

Directional Alternative

Example 9.4.9 Harvest Moon Festival

- Can people who are close to death postpone dying until after a symbolically meaningful occasion?
  - **H_0** : Given that an elderly Chinese woman dies within one week of the Harvest Moon Festival, she is equally likely to die before the festival or after the festival.
    Pr {die after festival} ≤ 1/2
  - **H_A** : Given that an elderly Chinese woman dies within one week of the Harvest Moon Festival, she is more likely to die after the festival than before the festival.
    Pr {die after festival} > 1/2
  - $\chi^2_s = (33 - 51.5)^2/51.5 + (70 - 51.5)^2/51.5 = 13.3$
  - $0.0001 < \text{P-value} < 0.001$ (non-directional Table 9)
  - $0.0001/2 < \text{P-value} < 0.001/2$ (directional Table 9)
  - We have strong evidence against H_0 and in favor of H_A
  - We conclude that the evidence is very strong that the death rate among elderly Chinese women goes up after the festival.

Table 9.4.4 Harvest moon festival data

	Before	After	Total
Observed	33	70	103
Expected	51.5	51.5	103

TABLE 9 Critical Values of the Chi-Square Distribution

Note: Column headings are non-directional (omni-directional) *P*-values. If H_A is directional (which is only possible when df = 1), the directional *P*-values are found by dividing the column headings in half.

	TAIL PROBABILITY						
df	0.20	0.10	0.05	0.02	0.01	0.001	0.0001
1	1.64	2.71	3.84	5.41	6.63	10.83	15.14

Summary

Chapter 9. Categorical Data: One-Sample Distribution

- 9.1 Dichotomous Observations
- 9.2 Confidence Interval for a Population Proportion
- 9.4 Inference for Proportions: The Chi-Square Goodness-of-Fit Test

Homework

Chapter 9

- 9.1.3
- 9.2.4; 9.2.9
- 9.4.3; 9.4.4