MATH 104 Data and Decisions Spring 2017

Study Sheet for Exam #3
Statistics

- This study sheet will not be allowed during the exam.
- Books, notes and online resources will not be allowed during the exam.
- A calculator will be handed out for use during the exam. **No other electronic devices (calculators, cell phones, tablets, laptops, etc.) will be allowed during the exam.**

Topics

1. Frequency tables and histograms
2. Mean, median, mode, range, variance, standard deviation
3. Basic probability
4. Continuous probability distributions
5. Normal distribution
6. Z-Scores and P-Values for the standard normal distribution
7. Z-Scores and P-Values for any normal distribution
8. Central Limit Theorem
9. Confidence intervals for the mean
10. Hypothesis testing for the mean

Practice Problems from Homework Exericses

- Exercise 3.1
- Exercise 3.2
- Exercise 3.3
- Exercise 3.4
- Exercise 3.5
- Exercise 3.7
- Exercise 3.8
- Exercise 3.9
- Exercise 3.10
- Exercise 3.12
- Exercise 3.13
- Exercise 3.14
- Exercise 3.15
- Exercise 3.16
- Exercise 3.17
- Exercise 3.18
- Exercise 3.19
- Exercise 3.20
- Exercise 3.21
- Exercise 3.22
- Exercise 3.23
- Exercise 3.24
- Exercise 3.25
- Exercise 3.26
- Exercise 3.27
- Exercise 3.31
- Exercise 3.32
- Exercise 3.33
- Exercise 3.34

Practice Problems from In-Class Exericses

- Exercise 3.4
- Exercise 3.5
- Exercise 3.6
- Exercise 3.7
- Exercise 3.8
- Exercise 3.9
- Exercise 3.11
- Exercise 3.12
- Exercise 3.14
- Exercise 3.15
- Exercise 3.16
- Exercise 3.17
- Exercise 3.18
- Exercise 3.19
- Exercise 3.20
- Exercise 3.21
- Exercise 3.22
- Exercise 3.23
- Exercise 3.24
- Exercise 3.25
- Exercise 3.26
- Exercise 3.27
- Exercise 3.28
- Exercise 3.31
- Exercise 3.32
- Exercise 3.33

Tips for Studying for the Exams

× **Bad** Forgetting about the homework.

✓ **Good** Making sure you know how to do all the relevant problems from the homework exercises and in-class exercises; seeking help from the instructor about the problems you do not know how to do.

× **Bad** Studying only by reading the texts.

✓ **Good** Doing exercises, and reading the texts as needed.

× **Bad** Studying only by yourself.

✓ **Good** Trying some exercises by yourself (or with friends), and then seeking help from the instructor about the exercises you do not know how to do.

× **Bad** Doing exercises while looking everything up in the texts or summaries.

✓ **Good** Doing some of the exercises the way you would do them on the exams, which is with closed book.

× **Bad** Staying up late (or all night) the night before the exam.

✓ **Good** Studying hard up through the day before the exam, but getting a good night's sleep the night before the exam.

Ethan's Office Hours

- **Monday:** 11:00-12:30
- **Tuesday:** 2:00-3:30 & 5:00-6:00
- **Thursday:** 2:30-4:00
- **Or by appointment**

Concepts and Formulas You Need to Know for the Exam

1. Statistics Terminology

1. A **population** is the complete collection of objects of interest.
2. A **parameter** of the population is a number that summarizes some aspect of the population.
3. A **sample** of the population is a subcollection of the population.
4. A **statistic** of the sample is a number that summarizes some aspect of the sample.

2. Frequency Table

Suppose that in a given set of numerical data, the values that appear (possibly more than once each) are $x_1, x_2, \ldots, x_n$. A **frequency table** for the data is a chart of the form

$$\begin{array}{c|cccc} x & x_1 & x_2 & \cdots & x_n \\ \hline f & f_1 & f_2 & \cdots & f_n \end{array},$$

where $f_1, f_2, \ldots, f_n$ are the number of times each of $x_1, x_2, \ldots, x_n$ occurs, in the data, respectively.

3. Mean

Let $x_1, x_2, x_3, \ldots, x_n$ be a collection of n numbers. The **mean** of these numbers, denoted $\bar{x}$, is defined by

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum x}{n} = \frac{1}{n} \sum x.$$

4. Mean from Frequency Table

Suppose that numerical data is given in a frequency table of the form

$$\begin{array}{c|cccc} x & x_1 & x_2 & \cdots & x_n \\ \hline f & f_1 & f_2 & \cdots & f_n \end{array}.$$

The mean of this data is computed by

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + f_3 x_3 + \cdots + f_n x_n}{f_1 + f_2 + f_3 + \cdots + f_n} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{\sum f x}{\sum f}.$$

5. Sample Mean vs. Population Mean

Suppose that a sample is taken from a population.

1. The **sample mean**, which is the mean of the sample, is denoted $\bar{x}$.
2. The **population mean**, which is the mean of the whole population, is denoted μ.

6. Median

Let $x_1, x_2, x_3, \dots, x_n$ be a collection of n numbers.

1. The **median** of these numbers, denoted $\tilde{x}$, is the number such that half of $x_1, x_2, x_3, \dots, x_n$ are above it and half of $x_1, x_2, x_3, \dots, x_n$ are below it.

2. To find the median, first list the numbers $x_1, x_2, x_3, \dots, x_n$ in increasing order.

   1. Suppose n is odd. Then the median is middle number, which is the entry x_i, where i is obtained by rounding the fraction $\frac{n}{2}$ up to the nearest whole number.
   2. Suppose n is even. Then the median is the average of the two middle numbers, which is $\frac{x_i + x_{i+1}}{2}$, where i is the whole number $\frac{n}{2}$.

7. Mode

Let $x_1, x_2, x_3, \dots, x_n$ be a collection of n numbers. The **mode** of these numbers is the value that occurs most frequently among the numbers.

8. Range

Let $x_1, x_2, x_3, \dots, x_n$ be a collection of n numbers. The **range** of these numbers is the difference between the largest value among these numbers and the smallest value among these numbers.

9. Sample Variance and Standard Deviation

Let $x_1, x_2, x_3, \dots, x_n$ be a sample.

1. The **sample variance** of these numbers, denoted s^2, is defined by

$$\begin{aligned} s^2 &= \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} \\ &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{\sum (x - \bar{x})^2}{n-1}. \end{aligned}$$

2. The **sample standard deviation** of these numbers, denoted s, is defined by

$$\begin{aligned} s &= \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}} \\ &= \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}. \end{aligned}$$

3. Note that the denominator of the above fractions is $n - 1$.

10. Sample Variance and Standard Deviation from Frequency Table

Suppose that numerical data is given in a frequency table of the form

$$\begin{array}{c|cccc} x & x_1 & x_2 & \cdots & x_n \\ \hline f & f_1 & f_2 & \cdots & f_n \end{array}.$$

1. The sample variance of this data is computed by

$$\begin{aligned} s^2 &= \frac{f_1(x_1-\bar{x})^2 + f_2(x_2-\bar{x})^2 + f_3(x_3-\bar{x})^2 + \cdots + f_n(x_n-\bar{x})^2}{f_1+f_2+f_3+\cdots+f_n-1} \\ &= \frac{\sum_{i=1}^n f_i(x_i-\bar{x})^2}{\left(\sum_{i=1}^n f_i\right)-1} = \frac{\sum f(x-\bar{x})^2}{\left(\sum f\right)-1}. \end{aligned}$$

2. The sample standard deviation of this data is computed by

$$\begin{aligned} s &= \sqrt{\frac{f_1(x_1-\bar{x})^2 + f_2(x_2-\bar{x})^2 + f_3(x_3-\bar{x})^2 + \cdots + f_n(x_n-\bar{x})^2}{f_1+f_2+f_3+\cdots+f_n-1}} \\ &= \sqrt{\frac{\sum_{i=1}^n f_i(x_i-\bar{x})^2}{\left(\sum_{i=1}^n f_i\right)-1}} = \sqrt{\frac{\sum f(x-\bar{x})^2}{\left(\sum f\right)-1}}. \end{aligned}$$

11. Probability Terminology

1. A **random variable** is a numerical variable the value of which depends upon a random phenomenon.

2. A **sample space** for a random variable is the set of all possible distinct outcomes for the random variable.

3. An **event** for a random variable is a subcollection of the sample space.

4. The **probability** of an event is a number between 0 and 1 (including those two numbers) that measures the likelihood of the event occurring.

5. If E is an event, the probability of E is denoted $P(E)$.

12. Discrete Probability

Suppose that a random variable has a finite sample space, in which all elements of the sample space are equally likely. Let E be an event for this random variable. Then

$$\begin{aligned} P(E) &= \frac{\text{number of elements in } E}{\text{number of elements in the sample space}} \\ &= \frac{\text{number of successful outcomes}}{\text{total number of possible outcomes}}. \end{aligned}$$

13. Probability of a Complement

Let S be a sample space and let E be an event. Then

$$P(\text{not } E) = 1 - P(E).$$

14. Probability of an Intersection

Let S be a sample space and let A and B be events. Suppose that A and B are independent. Then

$$P(A \text{ and } B) = P(A) \cdot P(B).$$

This formula is also written

$$P(A \cap B) = P(A) \cdot P(B).$$

15. Disjoint

Let A, B and S be sets. Suppose that A and B are subsets of S. The sets A and B are **disjoint** (also called **mutually exclusive**) if there are no elements of S that are in both A and B.

16. Probability of a Disjoint Union

Let S be a sample space and let A and B be events. Suppose that A and B are disjoint. Then

$$P(A \text{ or } B) = P(A) + P(B).$$

This formula is also written

$$P(A \cup B) = P(A) + P(B).$$

17. Probability Density Function

A **probability density function** (abbreviated **PDF**) is a function that allows us to compute the probability of a continuous random variable X.

A probability density function satisfies the following conditions.

1. The probability density function is continuous.
2. The probability density function is never negative.
3. The area under the whole probability density function and above the x-axis is 1.
4. If a and b are numbers such that $a < b$, then $P(a < X < b)$ is the area under the probability density function and above the x-axis between the values $x = a$ and $x = b$.

18. Normal Distribution

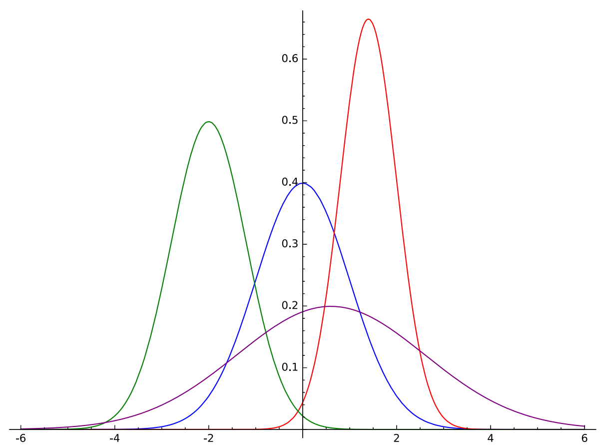
The **normal distribution** (also knows as the **Gaussian distribution**, or **bell-shaped curve**), is a type of probability density function that is very widely used in probability and statistics.

1. Each normal distribution is determined by two parameters, which are the mean, denoted μ, and the standard deviation, denoted σ. The value of μ can be any number, and the value of σ must be positive.

2. The normal distribution with mean μ and standard deviation σ is denoted $N(\mu, \sigma)$.

3. All normal distributions have similar shapes. The choice of μ moves the normal distribution to the right or left. The choice of σ makes the normal distribution either taller and thinner, or shorter and wider.

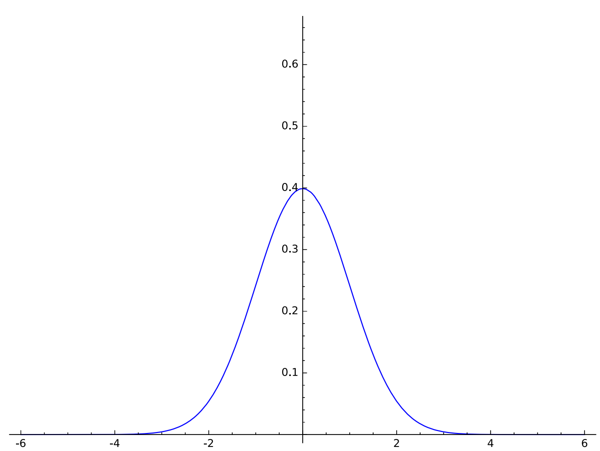
4. Some normal distributions are shown in below.



19. Standard Normal Distribution

1. The **standard normal distribution** is the normal distribution with $\mu = 0$ and $\sigma = 1$. The standard normal distribution is denoted $N(0, 1)$, also written $N(\mu = 0, \sigma = 1)$.

2. The graph of the standard normal distribution is shown below.



20. Z-Score Cutoff Points

Let c be a positive number less than 0.5.

1. The number $\mathbf{z_c}$ is defined to be the number such that $P\left(Z > z_c\right) = c$.
2. By symmetry, it follows that $P\left(Z < -z_c\right) = c$.
3. To find z_c, it is easier to find $-z_c$ first, and then obtain z_c by negating $-z_c$.

21. X-Value to Z-Score: Generic

Let X be a random variable. Suppose X follows a normal distribution with mean μ and standard deviation σ. This normal distribution is transformed to the standard normal distribution via the transformation

$$Z = \frac{X - \mu}{\sigma}.$$

The value of Z obtained in this way is called the Z-score for X.

22. Probability for X-Values

Let X be a random variable. Suppose X follows a normal distribution with mean μ and standard deviation σ. Let a and b be numbers. Suppose $a < b$.

1. $$P(X < b) = P\left(Z < \frac{b - \mu}{\sigma}\right).$$
2. $$P(a < X) = P\left(\frac{a - \mu}{\sigma} < Z\right) = 1 - P\left(Z < \frac{a - \mu}{\sigma}\right).$$
3. $$\begin{aligned} P(a < X < b) &= P\left(\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma}\right) \\ &= P\left(Z < \frac{b - \mu}{\sigma}\right) - P\left(Z < \frac{a - \mu}{\sigma}\right). \end{aligned}$$

23. Z-Score to X-Value: Generic

Let X be a random variable. Suppose X follows a normal distribution with mean μ and standard deviation σ. Let Z be a Z-score for the standard normal distribution. Then Z is transformed to the corresponding X-value via the transformation

$$X = \mu + Z\sigma.$$

24. Distribution of Averages

Let X be a random variable (with any type of distribution). Suppose X has mean μ and standard deviation σ.

Suppose that X is sampled n times, and the mean $\bar{X}$ is computed. If sampling and computing the mean is done repeatedly, it leads to a new distribution, called the **sampling distribution**, which is the distribution of the mean $\bar{X}$.

1. The mean of the sampling distribution is called the **sample mean** and is denoted $\mu_{\bar{X}}$.
2. The standard deviation of the sampling distribution is called the **sample standard deviation** (also known as the **standard error**) and is denoted $\sigma_{\bar{X}}$.

25. Central Limit Theorem

Let X be a random variable (with any type of distribution). Suppose X has mean μ and standard deviation σ.

The **Central Limit Theorem** states that for large sample sizes n (generally 30 or more), the sampling distribution is approximately a normal distribution, and that

$$\mu_{\bar{X}} = \mu \quad \text{and} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

26. Confidence Interval: Basic Idea

Let X be a random variable. Suppose a sample is taken, with sample mean $\bar{x}$.

1. The **level of confidence** (aka **confidence level**) for confidence intervals is a percentage strictly between 0% and 100%. The larger the percentage, the more confident we are in the result, though the harder it is to achieve that level of confidence. The level of confidence is specified by a positive number α that is less than 1, where the level of confidence equals $100(1-\alpha)\%$.
2. A **confidence interval** for the population mean μ at the $100(1-\alpha)\%$ level of confidence is an interval of the form $[\bar{x} - E, \bar{x} + E]$, for some number E, such that the probability that the interval $[\bar{x} - E, \bar{x} + E]$ actually contains the true value of μ is $100(1-\alpha)\%$.
3. The **margin of error** of a confidence interval of the form $[\bar{x} - E, \bar{x} + E]$ is the number E.

27. Confidence Intervals: Z-Scores

The Z-score confidence interval at the $100(1-\alpha)\%$ level of confidence is

$$\left[-z_{\alpha/2}, z_{\alpha/2}\right].$$

28. Confidence Intervals: Z-Score to X-Value

Let X be a random variable. Suppose X follows a normal distribution. Suppose a sample of size n is taken, with sample mean $\bar{x}$. Let Z be a Z-score for the standard normal distribution. For computing confidence intervals, the score Z is transformed to the corresponding X-value as follows.

1. If the population standard deviation σ is known, use the transformation

$$X = \bar{x} + Z \cdot \frac{\sigma}{\sqrt{n}}.$$

2. If the population standard deviation σ is not known, replace it is by the sample standard deviation s, and use the transformation

$$X = \bar{x} + Z \cdot \frac{s}{\sqrt{n}}.$$

29. Confidence Interval: General Formula

Let X be a random variable. Suppose X follows a normal distribution. Suppose a sample of size n is taken, with sample mean $\bar{x}$. The **confidence interval** for the population mean μ at the $100(1-\alpha)\%$ level of confidence is computed as follows.

1. If the population standard deviation σ is known, the confidence interval is

$$\left[\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right].$$

2. If the population standard deviation σ is not known, replace it is by the sample standard deviation s, and the confidence interval is

$$\left[\bar{x} - z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}\right].$$

30. Margin of Error

Let X be a random variable. Suppose X follows a normal distribution. Suppose a sample of size n is taken, with sample mean $\bar{x}$. The **margin of error** for the population mean μ at the $100(1-\alpha)\%$ level of confidence is computed as follows.

1. If the population standard deviation σ is known, the margin of error is
$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

2. If the population standard deviation σ is not known, replace it is by the sample standard deviation s, and the margin of error is
$$E = z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}.$$

3. Whether or not the population standard deviation σ is known, the confidence interval can be written
$$[\bar{x} - E, \bar{x} + E].$$

31. Hypothesis Testing: Basic Idea

Let X be a random variable. Suppose X follows a normal distribution. Suppose a sample of size n is taken, with sample mean $\bar{x}$.

1. The **level of significance** (aka **significance level**) for hypothesis testing is a percentage strictly between 0% and 100%. The smaller the percentage, the more confident we are in the result, though the harder it is to achieve that level of significance. The level of significance is specified by a positive number α that is less than 1, where the level of significance equals $100\alpha\%$.

2. The **null hypothesis**, denoted H_0, is the hypothesis we assume to be true unless there is sufficient evidence to reject it.

3. The **alternative hypothesis**, denoted H_a, is the hypothesis we would support if there is sufficient evidence, and only if there is sufficient evidence, to reject the null hypothesis.

4. A **hypothesis testing** for the sample mean $\bar{x}$ at the $100\alpha\%$ level of significance is a procedure to determine whether there is sufficient evidence to deduce that the sample mean $\bar{x}$ is sufficiently different from what the null hypothesis states that the null hypothesis should be rejected.

32. Three Cases of Hypothesis Testing

There are three cases of hypothesis testing for the mean, depending upon the form that the null hypothesis takes.

Let μ_0 be a number.

	Left tail	Two-sided	Right tail
Actual Null Hypothesis	$H_0: \mu_0 \leq \mu$	$H_0: \mu = \mu_0$	$H_0: \mu \leq \mu_0$
Used Null Hypothesis	$H_0: \mu = \mu_0$	$H_0: \mu = \mu_0$	$H_0: \mu = \mu_0$
Alternative Hypothesis	$H_a: \mu < \mu_0$	$H_a: \mu \neq \mu_0$	$H_a: \mu_0 < \mu$

The name of each of the three cases corresponds to the nature of where the region to reject the null hypothesis is located.

33. Hypothesis Testing: Z-Scores

Let μ_0 be a number. Let α be the level of significance for hypothesis testing.

The Z-scores for rejecting and not rejecting the null hypothesis are as follows.

Left tail

Null Hypothesis: $H_0: \mu_0 \leq \mu$

Alternative Hypothesis: $H_a: \mu < \mu_0$

Modified Confidence Interval (Non-Rejection): $\left(-\infty, z_\alpha\right]$

Rejection Region: $\left[z_\alpha, \infty\right)$.

Two-sided

Null Hypothesis: $H_0: \mu = \mu_0$

Alternative Hypothesis: $H_a: \mu \neq \mu_0$

Modified Confidence Interval (Non-Rejection): $\left[-z_{\alpha/2}, z_{\alpha/2}\right]$

Rejection Region: $\left(-\infty, -z_{\alpha/2}\right]$ & $\left[z_{\alpha/2}, \infty\right)$.

Right tail

Null Hypothesis: $H_0: \mu \leq \mu_0$

Alternative Hypothesis: $H_a: \mu_0 < \mu$

Modified Confidence Interval (Non-Rejection): $\left[-z_\alpha, \infty\right)$

Rejection Region: $\left(-\infty, -z_\alpha\right]$.

34. Hypothesis Testing: Z-Score to X-Value

Let X be a random variable. Suppose X follows a normal distribution. Suppose a sample of size n is taken, with sample mean $\bar{x}$. Let Z be a Z-score for the standard normal distribution. For hypothesis testing, the score Z is transformed to the corresponding X-value as follows.

1. If the population standard deviation σ is known, use the transformation

$$X = \bar{x} + Z \cdot \frac{\sigma}{\sqrt{n}}.$$

2. If the population standard deviation σ is not known, replace it is by the sample standard deviation s, and use the transformation

$$X = \bar{x} + Z \cdot \frac{s}{\sqrt{n}}.$$

35. Hypothesis Testing: X-Values

Let μ_0 be a number. Let α be the level of significance for hypothesis testing.

If μ_0 is in the modified confidence interval, the null hypothesis is not rejected, and if μ_0 is in the rejection region, the null hypothesis is rejected.

Left tail

Null Hypothesis: $H_0: \ \mu_0 \leq \mu$

Alternative Hypothesis: $H_a: \ \mu < \mu_0$

Modified Confidence Interval (Non-Rejection): $\left(-\infty, \bar{x} + z_\alpha \cdot \frac{s}{\sqrt{n}}\right]$

Rejection Region: $\left[\bar{x} + z_\alpha \cdot \frac{s}{\sqrt{n}}, \infty\right)$.

Two-sided

Null Hypothesis: $H_0: \ \mu = \mu_0$

Alternative Hypothesis: $H_a: \ \mu \neq \mu_0$

Modified Confidence Interval (Non-Rejection): $\left[\bar{x} - z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}\right]$

Rejection Region: $\left(-\infty, \bar{x} - z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}\right]$ & $\left[\bar{x} + z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}, \infty\right)$.

Right tail

Null Hypothesis: $H_0: \ \mu \leq \mu_0$

Alternative Hypothesis: $H_a: \ \mu_0 < \mu$

Modified Confidence Interval (Non-Rejection): $\left[\bar{x} - z_\alpha \cdot \frac{s}{\sqrt{n}}, \infty\right)$

Rejection Region: $\left(-\infty, \bar{x} - z_\alpha \cdot \frac{s}{\sqrt{n}}\right]$.

36. Hypothesis Testing: Errors

The possible errors in hypothesis testing are summarized in the following chart.

		Actually True or False	
		H_0 is true	H_0 is false
Our Decision	H_0 is not rejected	Good	Type 2 Error
	H_0 is rejected	Type 1 Error	Good

1. A **type 1 error** for hypothesis testing is when we reject the null hypothesis, even though it is actually true.

2. A **type 2 error** for hypothesis testing is when we do not reject the null hypothesis even though it is false.