STAT 5000

Statistical Methods I

Week 1

Fall 2024

Dr. Danica Ommen

Unit 0

Preliminaries

Who should take this class?

Difference between STAT 5000 and STAT 5870

- STAT 5870 (for graduate students outside of statistics)
- STAT 5000 (for graduate students major in statistics)

- Prerequisite for STAT 5000: STAT 5880 or current (or previous) enrollment in STAT 5420, knowledge of matrix algebra

Course Objectives

- Explore Procedures for Collecting Useful Data
- Display and Analyze Data
- Make Inferences about Populations
- Regression Analysis, Analysis of Variance, and other applications
- Obtain a Foundation in Linear Model Theory
- Become Familiar with SAS and R for Computing

Canvas Course WebPage

- Announcements
- Syllabus
- Modules (this is where you'll find course materials)
- Assignments
- Grades

Canvas Modules

Posted Weekly:

- Reading Assignments
- Lecture Slides
- Access to Homework, Labs, & Exams (including solutions)
- Data & SAS/R Code

Assessment

Grade Break-down

- Homework: 20%
- Labs: 20%
- Exams: 60% (3 exams each worth 20%)

- Late work is not accepted after the solutions are posted!
- Lowest Homework & Lab dropped at end of semester

Computing Resources

- Scientific Calculator
- SAS (free student account available)
- R (free for all)
- Student-owned laptop to run SAS & R

University & Other Course Policies

Posted in Canvas Syllabus:

- Textbooks
- Free Expression
- Academic Dishonesty
- Accessibility
- Discrimination and Harassment
- Mental Health and Wellbeing
- Prep Week
- Religious Accommodation
- Course Summary

Questions?

Contact me:

dmommen@iastate.edu

Visit Student Office Hours - TBD

Unit 1

Introduction

What is Statistics?

Dictionary Definitions:

- Branch of mathematics dealing with the collection, analysis, interpretation and presentation of data
- Art and science of drawing justifiable conclusions from data

Statistics as Mathematics

- Simple Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

  where $\epsilon_i \sim$ iid $N(0, \sigma^2)$, and $i = 1, 2, \cdots, n$.
- Model parameters are β_0, β_1 and σ^2
- We will find estimators for these parameters and derive their properties.

Statistics as Mathematics

- The simple linear regression model in matrix form:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \text{ where } \epsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

- The matrix formulation has

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \text{ and } E(\mathbf{Y}) = \mathbf{X}\beta = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

- The unknown parameters are $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$ and σ^2

Statistics as Mathematics

We have the following results:

- The least squares estimator

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$$

  is the minimum variance linear unbiased estimator for β
- $\text{Var}(\hat{\beta}) = \mathbf{V} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$
- $\mathbf{c}^T\hat{\beta} \sim N(\mathbf{c}^T\beta,\ \mathbf{c}^T\mathbf{V}\mathbf{c})$
- Test H_0: $\mathbf{c}^T\beta = 0$ using $t = \dfrac{\mathbf{c}^T\hat{\beta} - 0}{\sqrt{\mathbf{c}^T\mathbf{V}\mathbf{c}}}$

Statistics as Art

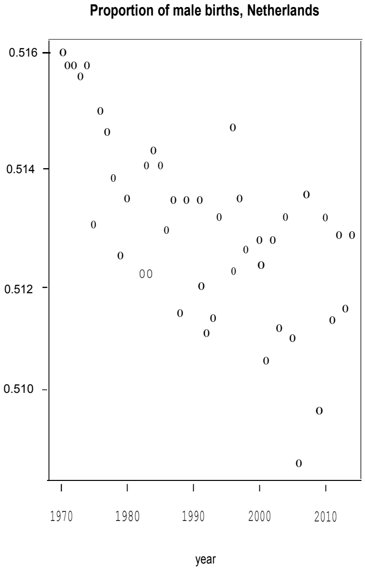
Suppose a researcher says to you:

I have data on percentages of male births in the Netherlands from 1970 to 2013. They seem to be declining. I need to forecast the expected change in the next ten years (2014 - 2023). You're a statistician; can you help me?

Some issues:

- What type of model can be used to address the question?
- How can you determine if a proposed model is reasonable?

Statistics as Art



Statistics as Art

What types of inference can be made?

- Did a significant decline in male births occur?
- Will trends continue into the future?
- What types of predictions can be made?
- How can accuracy or reliability of predictions be assessed?

- Usually more than one reasonable approach.
- Models are almost never "true", but some may be useful?

Statistics as Science

Statistics is the science of using information to make decisions and quantify uncertainty inherent to those decisions.

There are four basic steps in the statistical problem solving process (Deming):

1. Define the questions to be answered (Plan)
2. Gather appropriate data (Do)
3. Analyze the data (Study)
4. Interpret the results (Act)

Define Questions (Plan)

- Researchers define study questions.
- As statisticians, sometimes we play a role in question development, sometimes not.
- Questions drive rest of statistical investigation.

Define Questions (Plan)

Examples:

- Industrial experiments
  - improve process yields and quality?
  - reduce production costs and increase profits?
  - reduce sensitivity to uncontrolled factors (variation)?
- Agriculture and food sciences
  - improve crop yields?
  - improve disease resistance and manage pests?
  - develop new food products?
- Human health studies
  - establish effectiveness of treatments
  - identify causes of disease
  - assess the role of nutrition in long term health status

Define Questions (Plan)

Examples:

- Business and economics
  - model consumer behavior?
  - inventory control?
  - improve management processes (business analytics)?
- Ecology
  - Monitor population growth?
  - Examine competition among species?
  - Monitor water or air quality?
  - Examine climate change?

Statistics as Science

Data Collection (Do)

Experiments: Researchers impose an intervention on members of some population

- Planned intervention (prospective)
  - Researcher changes the level of at least one factor in order to observe a response
- Causal inferences are possible
  - Hold the levels of all other factors constant
  - Random assignment of experimental units to treatment groups(randomized experiment)
    - Randomization eliminates uncontrolled sources of bias
    - Randomization provides a basis for inference

Statistics as Science

Data Collection (Do)

Observational studies: Members of some population are observed as they naturally exist

- Census: Observe all members of some population
- Haphazard sample
- Representative random sample
- May be able to make inferences about associations, but causal inferences are usually not be possible.

Statistics as Science

Data Analysis (Study)

- Methods depend on:
  - Type of data collected
  - How data were collected
  - Research Questions
- Typically what outsiders consider **Statistics**

Interpret Results (Act)

- Answer research questions based on results from data analysis
- Conclusions must match scope of data collection methods
  - Random Assignment of Units to Groups = Experiment
    - Causal inferences can be drawn
  - No Random Assignment = Observational study
    - No causal inference
  - Random Selection of Units in Sample (experiment or observational study)
    - Generalization of inference to population possible

Statistics is ...

1. Part Mathematics
2. Part Art
3. Part Science

STAT 5000 will include all three aspects

Unit 1

Experiments

Terminology

Experiment: an investigation in which the investigator applies (assigns) some treatments to experimental units and then observes the effect of the treatments on the experimental units by measuring one or more response variables.

Treatment: a condition or set of conditions applied to experimental units in an experiment.

Experimental Design: The assignment rule specifies which experimental units are to be observed under which treatments.

Experiments

Terminology

Experimental Unit: the physical entity to which a treatment is randomly assigned and independently applied.

- the smallest division of material (e.g. land, plant, animal, etc) to be studied

Response Variable: a characteristic of an experimental unit that is measured after treatment and analyzed to assess the effects of treatments on experimental units. (e.g. yield, gene expression level, etc.)

Observational Unit: the unit on which a response variable is measured. There is often a one-to-one correspondence between experimental units and observational units, but that is not always true.

Basic Principles

Replication

- Applying a treatment independently to two or more experimental units
- Level of variability can be estimated for units that are treated alike.

Randomization

Experimental Design

Basic Principles

Replication

Randomization

- Random assignment of treatments to experimental units
- Reduce or eliminate sources of bias (treatment groups are equivalent, *on average*, except for the assigned treatment)
- Cause and effect relationships can be demonstrated
- Create a probability distribution for a test statistic under the null hypothesis of no treatment effects

Experimental Design

Basic Principles

Blocking/Matching

- Group similar experimental units into blocks
- Apply each treatment to (the same number of) experimental units within each block (balance)
- Separate random assignment of units to treatments is done within each block (randomization)

Blinding

- Subjects do not know which treatment they received
- Researchers making measurements do not know the treatment assignments

Experimental Design

Basic Principles

Control of Extraneous Variables

- Control non-intervention factors
- Use homogeneous experimental units
- Accurate measurement of outcomes (responses)
- Tradeoff between accuracy and generalizability

Comparison to a Control Group

- Untreated (placebo) group
- Gold standard (best available treatment)

Basic Principles

- Inferences are restricted to only those units used in the experiment
- Extending inferences beyond the units in the experiment
  - Were the units used in the experiment obtained from **a representative random sample** from some larger population?
    - Yes $\Rightarrow$ can make inferences about the population
    - No $\Rightarrow$ cannot make inferences about the population

Study Ethics

- Does the potential gain from performing (or continuing) the study outweigh the risks to the subjects?
  - Some experiments are unethical
  - Human and animal safety committees
  - Data Safety Monitoring Boards
- Informed Consent
  - Subjects must be informed of objectives and risks
  - Subjects must agree to participate
  - Subjects must be free to withdraw at any time (need to deal with incomplete sets of responses)

Planning and Execution

Idealized story

- Clearly define objectives
- Identify population of interest
- Obtain a random sample of units for the study
- Random assignment of units to treatment groups
- Apply treatments to units and record responses
- Analyze the data
- Report inferences about the population

Planning and Execution

Reality

- Objectives are not always clear
- Non-random samples of convenience
  - Volunteers
  - Subjects seeking treatment at a medical clinic
- Cannot generalize inference to population

Example: Low Birthweight Babies

Objective:

- Determine if a specific program of assistance to families of low birthweight babies can increase weight gain during the first 12 months after release from the hospital?

Preliminary Planning:

- Gather expert information
  - Identify factors that could affect weight gain
  - Which factors can be controlled?
  - What outcomes will be measured?
- Identify source of subjects (low birthweight babies)

Example: Low Birthweight Babies

Develop a Plan (Protocol):

- Identify restrictions on recruitment
- Use of blocking
- Use of randomization
- Use of blinding
- Describe how and when responses will be measured
- Specify methods of analysis
- Determine sample sizes and budget
- Obtain ethical approval

Example: Low Birthweight Babies

Implement the Plan:

- Secure adequate resources
- Perform the experiment
- Analyze the data
- Interpret results
- Report conclusions and recommendations

Example: Low Birthweight Babies

Experiment Details:

- Two treatments (two levels of one factor)
  - Printed information only
  - Printed information and scheduled visits by a nurse
- Experimental units:
  - low birthweight babies born in a set of participating hospitals (and their families)
- Blocking factor:
  - Three weight classes

Example: Low Birthweight Babies

Experimental Details:

- Replication:
  - 30 babies in each weight class
- Randomization:
  - random assignment of 15 babies to each treatment within each weight class
- Control of extraneous variation:
  - Babies with certain birth defects and illnesses were excluded
- Measured weight gain during the first year after entering study

Example: 1954 Salk Polio Vaccine Study

Design 1: National Foundation for Infantile Paralysis (NFIP) (“observed control", not randomized)

- grades 1 and 3 $\Rightarrow$ control
- grade 2 with consent $\Rightarrow$ vaccine
- grade 2 without consent $\Rightarrow$ control

- Problems with NFIP
  - Are grades 1 and 3 valid controls?
  - No consent group from grade 2 is not a valid control (different types of children)

Example: 1954 Salk Polio Vaccine Study

Design 2: Randomized clinical trial

- Ask parents of second grade children for consent to enroll their child in the study
  - randomly assign 1/2 to vaccine
  - randomly assign 1/2 to placebo
- No consent $\Rightarrow$ excluded from study
- Double blind study
  - Doctors did not know assignment
  - Children and parents did not know assignment

Example: 1954 Salk Polio Vaccine Study

Results:

	Randomized Trial	
Group	sample size	cases per 100,000
Vaccine	200K	28
Control	200K	71
No consent	350K	46

	NFIP study	
Group	sample size	cases per 100,000
Grade 2 (vaccine)	225K	25
Grade 1,3 (control)	725K	54
Grade 2 (no consent)	125K	44

Comparing the two tables:

- lines 1 and 3 are similar, but line 2's are different

Example: 1954 Salk Polio Vaccine Study

Randomization as a basis for inference

- In the randomized trial:
  56 children in the treatment group developed polio
  142 children in the control group developed polio
- If the vaccine had no effect, then a child would experience the same outcome in either group (198 children would develop polio)
- Of those 198 cases, 142 occurred in the control group
- Because we randomized, we can compute the probability that at least 142 of the 198 cases would occur by chance in the control group if the vaccine had no effect (1 in 2 billion)
- “proves” vaccine is effective

Unit 1

Randomization Tests

Randomization Tests

Scenario

- Randomized Experiment
  - Treatment variable: one factor with two levels
- Randomly assign experimental units to one of two treatment groups.

Research Question

- Is there a difference in the value of the response variable between the two treatments?
- Source of inference:
  Random assignment of experimental units to treatments

Notation

Parameters:

- $\mu_1 =$ mean response for Treatment 1
- $\sigma_1^2 =$ variance of response for Treatment 1
- $\sigma_1 =$ std. dev. of response for Treatment 1

- $\mu_2 =$ mean response for Treatment 2
- $\sigma_2^2 =$ variance of response for Treatment 2
- $\sigma_2 =$ std. dev. of response for Treatment 2

Notation

Data:

- $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$

  value of response variable for n_1 experimental units receiving treatment 1.

- $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$

  value of response variable for n_2 experimental units receiving treatment 2.

Notation

Summary Statistics

- Treatment 1

$$\overline{Y}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} Y_{1j}$$

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (Y_{1j} - \overline{Y}_1)^2, \qquad S_1 = \sqrt{S_1^2}$$

- Treatment 2

$$\overline{Y}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} Y_{2j}$$

$$S_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (Y_{2j} - \overline{Y}_2)^2, \qquad S_2 = \sqrt{S_2^2}$$

Methods of Analysis

- Reach conclusions and make recommendations using:
  - Visual displays
  - Point estimation: estimate $\mu_1,\ \mu_2,\ \sigma_1, \sigma_2,\ \mu_1 - \mu_2,$ etc.
  - Interval estimation: confidence intervals for $\mu_1 - \mu_2$, etc.
  - Tests of hypotheses ($\mu_1 = \mu_2$?)
- Types of inference
  - Randomization (design-based)
  - Model-based (relies on the specification of a model)

Basis for Inference

- Used for randomized experiments
- Use the probability distribution imposed by the **random** assignment of units to treatment groups
  - Under the null hypothesis
    H_0 : *treatments have the same effect*
    the response provided by any particular unit does not depend on the assigned treatment ($\Rightarrow \mu_1 = \mu_2$)
  - Is the observed difference $\bar{y}_1 - \bar{y}_2$ inconsistent with H_0?
  - Compare $\bar{y}_1 - \bar{y}_2$ with differences in sample means for all other possible random assignments of units to treatment groups (What if H_0 is true?)

Motivating Example: Rats Running

- Suppose we want to test whether a drug affects the running ability of rats.
- We randomly divide a group of eight rats into two groups of four.
  - Each rat in one group is injected with the drug.
  - Each rat in the other group is injected with a control substance.
- Then the running time before rest (in minutes) is measured for each rat.

Randomization Test

Motivating Example: Rats Running

Running Time in minutes (Hypothetical Data)

Control:	9	12	14	17
Drug:	18	21	23	26

- The average running time is 13 for the control group, and 22 for the drug group.
- Is this difference caused by the drug?

Randomization Test

Motivating Example: Rats Running

- Clearly there is some natural variation in the response variable (not due to treatment) because the running times differ among rats **within each treatment group**.
- Maybe the observed difference (22-13=9) showed up simply because it happened that the rats with better endurance were chosen for injection with the drug.
- What is the chance of seeing such a large difference in treatment means if the drug has no effect?

Randomization Test

Motivating Example: Rats Running

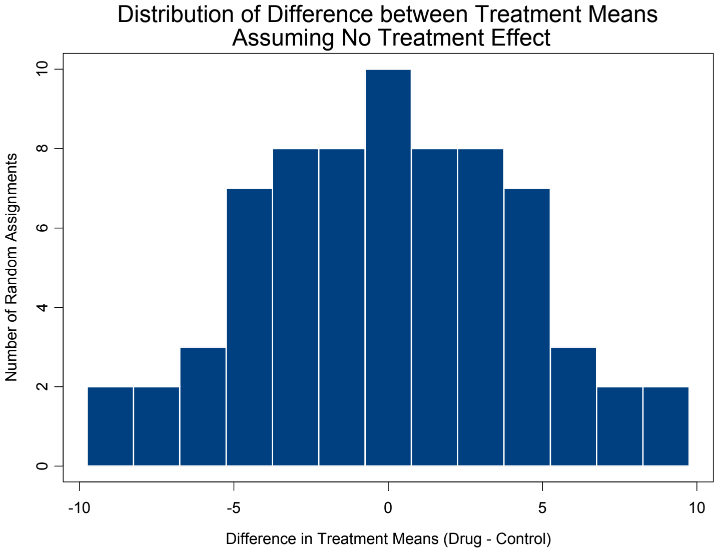
Let's perform a **randomization test** to see ...

- If the drug has no effect, then the treatments don't matter
- Thus, we can randomly reassign rats to treatment groups
- Then, re-compute the difference in means for the new treatment groups
- Do this a bunch of times
- Assess the results graphically and numerically
  - Graphically: randomization histogram
  - Numerically: randomization p-value

Randomization Test

Random									Difference
Assignment	Control				Drug				in Averages
1	9	12	14	17	18	21	23	26	9.0
2	9	12	14	18	17	21	23	26	8.5
3	9	12	14	21	17	18	23	26	7.0
4	9	12	14	23	17	18	21	26	6.0
5	9	12	14	26	17	18	21	23	4.5
6	9	12	17	18	14	21	23	26	7.0
7	9	12	17	21	14	18	23	26	5.5
8	9	12	17	23	14	18	21	26	4.5
9	9	12	17	26	14	18	21	23	3.0
10	9	12	18	21	14	17	23	26	5.0
11	9	12	18	23	14	17	21	26	4.0
12	9	12	18	26	14	17	21	23	2.5
13	9	12	21	23	14	17	18	26	2.5
14	9	12	21	26	14	17	18	23	1.0
15	9	12	23	26	14	17	18	21	0.0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
69	18	21	23	26	9	12	14	17	-8.5
70	18	21	23	26	9	12	14	17	-9.0

Randomization Test



Motivating Example: Rats Running

- Only 2 of the 70 possible random assignments would have led to a difference between treatment means as large as 9.
- Thus, under the assumption of no drug effect, the chance of seeing a difference as large as we observed was 2/70 = 0.0286.
- Because 0.0286 is a small probability, we have reason to attribute the observed difference to the effect of the drug rather than a coincidence due to the way we assigned our experimental units to treatment groups.

Example: Creative Writing

The Statistical Sleuth, Section 1.1:
T. Amabile, *J. Per. and Soc.l Psych.*, 48(2), 1985, 393-99

- Experimental units: experienced creative writers
- Treatments: questionnaires on motivation for writing given at the beginning of the study (SS, page 3)
  - intrinsic motivation (enjoyment, satisfaction, etc...)
  - extrinsic motivation (jobs, financial rewards, etc.)
- Random assignment: (24 intrinsic, 23 extrinsic)
- Response: Creativity displayed in writing a Haiku style poem on laughter $\Rightarrow$ average of evaluations by 12 poets on a 40 point scale

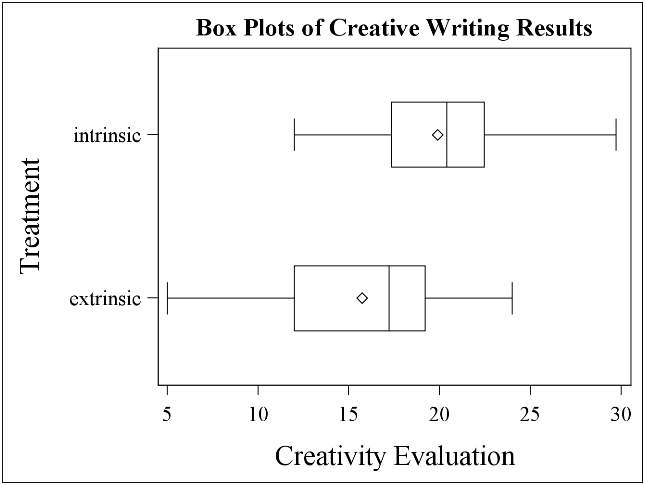
Example: Creative Writing

Observed data

intrinsic:	12.0	12.0	12.9	13.6	16.6	17.2
	17.5	18.2	19.1	19.3	19.8	20.3
	20.5	20.6	21.3	21.6	22.1	22.2
	22.6	23.1	24.0	24.3	26.7	29.7
extrinsic:	5.0	5.4	6.1	10.9	11.8	12.0
	12.3	14.8	15.0	16.8	17.2	17.2
	17.4	17.5	18.5	18.7	18.7	19.2
	19.5	20.7	21.2	22.1	24.0	

Example: Creative Writing

Data display



Example: Creative Writing

- Five summary statistics:

Treatment	min	Q1	median	Q3	max
1	12.0	17.35	20.40	22.45	29.70
2	5.0	12.00	17.20	19.20	24.00

- Sample means and standard deviations

$$\bar{y}_1 = 19.8875, \quad s_1 = 4.4418$$

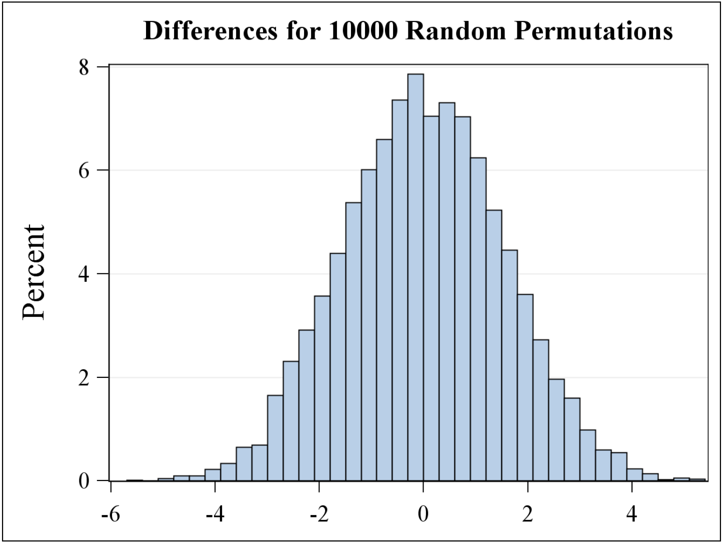
$$\bar{y}_2 = 15.7391, \quad s_2 = 5.2526$$

- Observed difference in sample means is 4.1484

Example: Creative Writing

- H_0 : Treatments 1 and 2 have the same effect on creativity
- H_a : Treatments 1 and 2 have different effects on creativity
- 1.6×10^{13} possible random assignments
- Sample of 10000 randomization assignments of subjects to treatments (assume the null hypothesis is true)
  - 50 of 10000 randomizations have values as large as 4.1484 or as small as -4.1484
  - extremely unlikely to see a difference this big by chance (two-tailed p-value = .0050)

Example: Creative Writing



Example: Creative Writing

Conclusions:

- questionnaire on intrinsic rewards leads to more creative writing in these students
- not a random sample … can’t necessarily infer that this is true in a larger population

Wrap-Up

General Comments

- The randomization test is also called the permutation test
- The randomization test (permutation test) depends on identifying units to permute, which should be the units in the experiment that are **exchangeable under the null hypothesis**, determined by the design of the experiment and the factor(s) being tested.