Review

EES 4891/5891

Probability & Statistics for Geosciences

Jonathan Gilligan

Class #17: Tuesday, March 18 2025

Announcement

Announcement

- Office Hours canceled this Wednesday
  - Make-Up Office Hours:
    - Thursday March 20, 2:00–3:30
    - Or by arrangement (Most of Friday is open)
- This week, I will be recording lectures to make up for missed classes.

Learning Goals

Learning Goals

- Refresh your understanding of:
  - Descriptive statistics
    - Learn the difference between *Pearson* and *Spearman* methods for estimating *correlation*.
  - The central limit theorem
  - How to estimate parameters for parametric distributions
  - Testing hypotheses
    - New material: Bayesian hypothesis testing

Semester Research Project

Semester Research Project

- Posted to course web site
- By March 27, choose a data set (from your own research, published data sources, etc.)
  - Does not need to be geological or environmental
  - Come see me, or email for help finding a source
  - The assignment lists many possible sources of public data
- Elements:
  1. Import your data into R
  2. Characterize your data:
     - Descriptive statistics
     - Plot distributions
     - Figure out what kind of distribution your data has,
     - …
- Elements (cont.):
  3. Develop questions about your data
     - Is there a relationship between different variables?
     - Are data from two different places or times similar or different?
     - …
  4. Use hypothesis tests for answering your questions
  5. Write up a report, using R and Quarto
     - You can write up sections of your report as you go through 1–4
- By Tues. Apr. 22, turn in your project:
  - Render Quarto to HTML, PDF, or Word
  - Push your project to GitHub Classroom

Review of Basics

Descriptive Statistics

- From a distribution $X \sim \mathcal{D}$:
  - mean $\mu = E(\mathcal{D})$
  - variance

    $$\text{Var}(\mathcal{D}) = \sigma^2 = E((X - E(X))^2)$$

  - standard deviation $\sigma = \sqrt{\text{Var}}$
- Estimates from a sample of n observations $x_1, x_2, \ldots, x_n$:
  - mean

    $$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - variance

    $$\hat{\sigma}^2 = s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

  - standard deviation

    $$\hat{\sigma} = s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Comparing Two Variables

- Covariance:

$$\begin{aligned}\mathrm{Cov}(X, Y) =& E((X - E(X))(Y - E(Y))) \\ =& E(XY) - E(X)E(Y)\end{aligned}$$

  - Covariance is symmetric: $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
  - $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$

- Correlation: Covariance has units of X and Y, which may be nonsensical ($^\circ\mathrm{C} \times p\mathrm{H}$)
  - Correlation is dimensionless (no units) and ranges from −1 to 1.

- Pearson correlation:

$$\rho_{X,Y} = \mathrm{Cov}\left(\frac{X - E(X)}{\sigma_X}\ \frac{Y - E(Y)}{\sigma_Y}\right)$$

- Pearson correlation measures linear correlation: if $\rho = \pm 1$, then X and Y have a perfect linear relationship.
- Spearman correlation measures nonlinear correlation:
  - Rank X and Y from smallest to largest (1 to n)
  - Spearman correlation ρ_{Spearman} is the Pearson correlation of the ranks.

Example

- Generate data

```
set.seed(12345)
x <- seq(0, 1, 0.01)
y <- sinh((x - 0.5) * 10)
```

  - Covariance:

```
cov(x, y)
```

```
## [1] 6.315366
```

  - Pearson correlation:

```
cor(x, y, method = "pearson")
```

```
## [1] 0.8748097
```

    - Pearson correlation $= \pm 1$ if x and y lie exactly on a line.

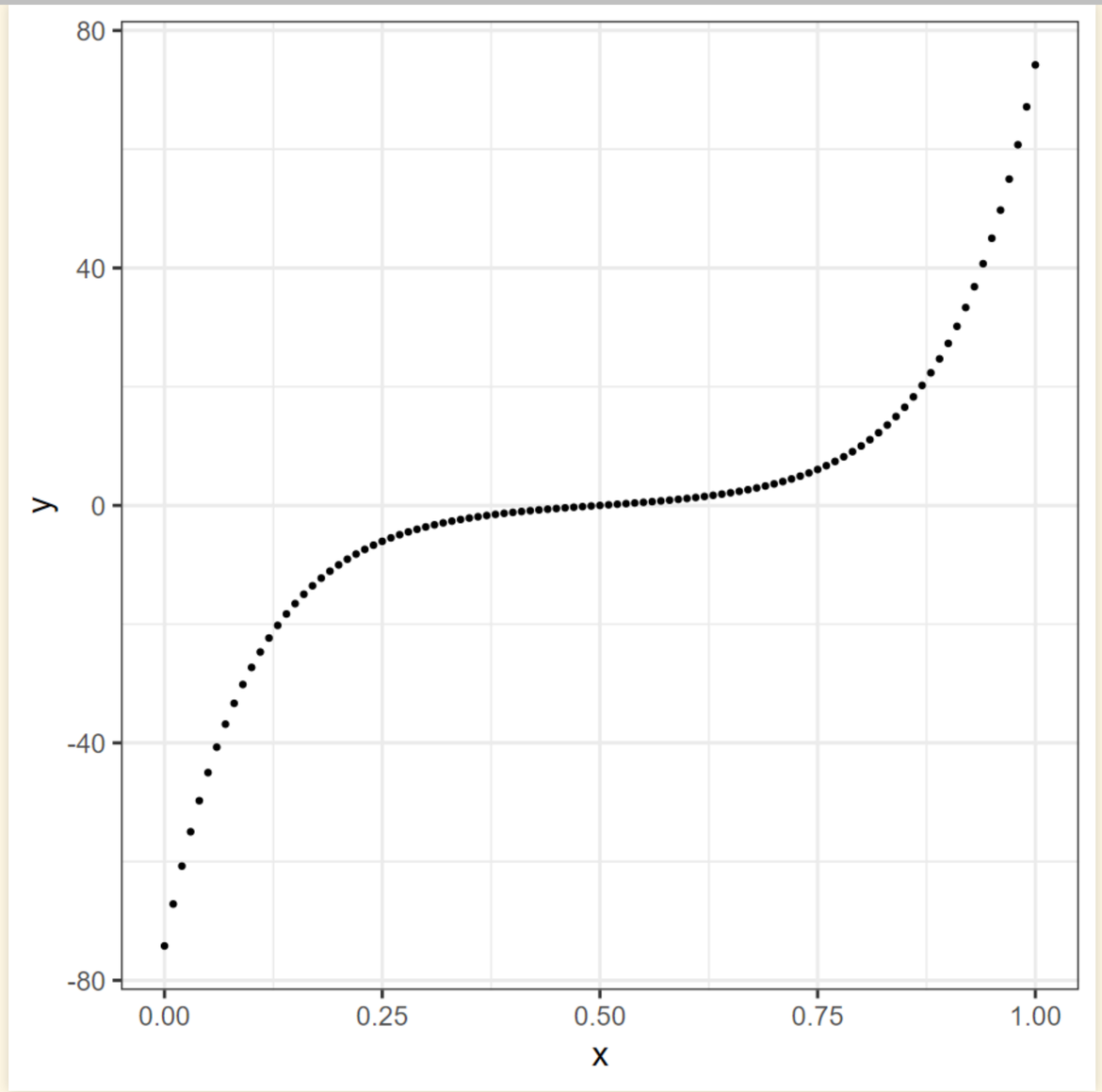
  - Spearman correlation:

```
cor(x, y, method = "spearman")
```

```
## [1] 1
```

    - Spearman correlation $= \pm 1$ if x and y lie exactly on any monotonic curve.

```
df <- tibble(x = x, y = y)
ggplot(df, aes(x = x, y = y)) + geom_point()
```



Central Limit Theorem

Central Limit Theorem

- Consider a set of M experiments, each taking n samples of a variable x, with $X \sim \mathcal{D}$, for some distribution $\mathcal{D}$
- For each experiment, $\overline{x}$ is the mean of the n observations of x from that experiment
- As $n \to \infty$, the distribution of $\overline{x}$ approaches a normal distribution with
  - Mean μ approaching $E(\mathcal{D})$
  - Variance σ approaching

$$\frac{1}{n}\text{Var}(\mathcal{D})$$

- No matter what distribution your data comes from,
  - If you make n observations, and n is large,
    - the *average* of the observations will be normally distributed.
    - The bigger n is the closer $\overline{x}$ will be to the true mean of the distribution.
- If you want to compare two experiments, each making n observations,
  - you can compare the averages, using tests that assume normality.
- But this only tells you about the averages of your data.
  - Often, you want to know about the distribution your data came from.
    - e.g., estimating parameters of that distribution.

Estimating Parameters

Estimating Parameters for a Normal Distribution

- For a normal distribution,
  - Estimate mean:

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - Estimate variance:

$$\hat{\sigma}^2 = s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i = \bar{x})^2$$

Estimating Parameters for Other Distributions

- Maximum-Likelihood Estimation
  - For data $X = (x_1, x_2, \ldots, x_n)$, and a distribution $\mathcal{D}(\Theta)$ with parameters $\Theta = (\theta_1, \theta_2, \ldots, \theta_p)$ and likelihood function $L(x|\Theta)$
    - Find values for Θ that maximize the joint likelihood of the observed x_i:

$$L(x_1, x_2, \ldots, x_n|\Theta) = \prod_{i=1}^{n} L(x_i|\Theta)$$

$$\log L(x_1, x_2, \ldots, x_n|\Theta) = \sum_{i=1}^{n} \log L(x_i|\Theta)$$

  - Using R to estimate parameters for a gamma distribution from observations x:

```
library(MASS)
fitdistr(x, "gamma")
```

- Bayesian Estimation:
  - Figure out *prior probability distributions* for $\theta_1, \theta_2, \ldots, \theta_p$
    - What did you already know about Θ before making the observations?
  - Use Bayes's theorem to calculate *posterior probability distributions* for $\theta_1, \theta_2, \ldots, \theta_p$

$$\underbrace{P(\Theta|X)}_{\text{posterior}} = \frac{\overbrace{L(X|\Theta)P(\Theta)}^{\text{likelihood}\quad\text{prior}}}{\underbrace{P(X)}_{\text{evidence}}}$$

$$= \frac{L(X|\Theta)P(\Theta)}{\int L(X|\Theta)P(\Theta)\mathrm{d}\Theta}$$

    - Challenge: $\int L(X|\Theta)P(\Theta)\mathrm{d}\Theta$ is hard to compute for more than a few parameters.

Testing Hypotheses

General Strategy

1. Identify the test and test statistic
   - e.g., *t*-test
2. Define the null hypothesis
   - e.g., $H_0 : \mu = \mu_0$
3. Define an alternate hypothesis
   - e.g., $H_a : \mu > \mu_0$
4. Calculate the null distribution
   - Test statistic, assuming H_0 is true
5. Calculate the *p*-value
   - Cumulative probability of the test statistic, assuming H_0 is true

- Test errors:
  - Type-I: H_0 is true, but you reject it (*false positive*)
    - You don't have a disease (H_0), but a clinical test says you do.
    - *p*-value is the probability of a Type-I error, if H_0 is true
  - Type-II: H_0 is false, but you don't reject it (*false negative*)
    - You have a disease (H_0 is false) but a clinical test says you don't.

Testing Hypotheses

- Kinds of tests:
  - What kind of distribution is it?
    - Q-Q Tests (any kind of distribution)
    - Shapiro-Wilk test (Normal only) (`shapiro.test(x)`)
  - Goodness of fit: Does the data match a distribution with specific parameter values
    - Chi-squared test `chisq.test()`
      - Mostly for discrete distributions: Binomial, Poisson, etc.
    - Kolmogorov-Smirnov test `ks.test()`
      - Only for continuous distributions
      - *p*-values are not reliable if you estimate parameters from the data.
        - Anderson-Darling tests are better if you estimate parameters from data: `AndersonDarlingTest()` from `DescTools` package.
- Testing fit parameters:
  - Compare observations to specific parameters: one-sample tessts
  - Compare two sets of observations (do they come from the same distribution?): two-sample tests
  - Normally distributed data:
    - If you know the variance *a priori* and want to compare means: Z-test
    - If you don't know the variance and want to compare means: t-test
    - If you want to compare variance: F-test

Nonparametric Tests

- Most parametric tests only work if your data follow a parametric distribution
  - Most work only work for a Normal distribution
  - This is a historical artifact of what math people could do using pencil and paper
  - For testing means, the Central Limit Theorem makes normal distributions widely applicable.
- Nonparametric tests substitute computer power for mathematical elegance
  - They work for any distribution
  - You don’t need to know a mathematical formula for the distribution
- Basic strategy:
  - Simulate a large sample of *surrogate data* under the null hypothesis
  - Compare this sample to the observed data
- Four ideas that use Monte Carlo methods:
  1. Permutation
  2. Reordering
  3. Resampling
  4. Direct simulation
- Monte-Carlo methods are also used for Bayesian analysis

Bayesian Hypothesis Testing

Bayesian Hypothesis Testing

- The problem:
  - The techniques we've learned only examine evidence that H_0 is *false*.
    - What about evidence that it's *true*?
- Bayesian methods compare a null hypothesis H_0 to an alternate hypothesis H_a and choose the one that has more evidence supporting it.
  - Bayes's Theorem: For a hypothesis H,

$$\underbrace{P(H|D)}_{\text{posterior}} = \frac{\overbrace{P(D|H)}^{\text{likelihood}}\ \overbrace{P(H)}^{\text{prior}}}{\underbrace{P(D)}_{\text{evidence}}}$$

where D is the new data and $P(H)$ is the prior probability that H is true, based on what you knew before observing D.

- Consider two hypotheses, H_1 and H_2:
  - The *Bayes factor* K is the ratio of probabilities that each is true:

$$K = \frac{P(H_1|D)}{P(H_2|D)} = \frac{\left(\frac{P(D|H_1)\ P(H_1)}{P(D)}\right)}{\left(\frac{P(D|H_2)\ P(H_2)}{P(D)}\right)}$$

  - $P(D)$ cancels out, so

$$K = \frac{P(D|H_1)\ P(H_1)}{P(D|H_2)\ P(H_2)}$$

  - If you had no reason to prefer H_1 or H_2, then $P(H_1)$ and $P(H_2)$ cancel out:

$$K = \frac{P(D|H_1)}{P(D|H_2)}$$

Bayesian Hypothesis Testing

- The *Bayes factor* K is the ratio of the posterior probabilities of H_1 and H_2 being true:

$$K = \frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)\,P(H_1)}{P(D|H_2)\,P(H_2)}$$

- If $K = 1$, H_1 and H_2 are equally likely.
- If $K > 1$, bigger K gives more confidence in H_1 vs. H_2
- If $K < 1$, the closer K is to 0, more confidence in H_2 vs. H_1
- This says nothing about whether a third hypothesis H_3 might be more likely than H_1 or H_2.

- There are more sophisticated Bayesian methods of comparing hypotheses.
  - Information Criteria
    - Bayesian Information Criterion (BIC) (avoid this)
    - Akaike Information Criterion (AIC) (good)
    - Watanabe-Akaike Information Criterion (WAIC) (better)
  - All of these use information theory to evaluate evidence, while accounting for degrees of freedom used in estimating parameters.
    - They allow you to compare many hypotheses all at once.
  - Beyond the scope of this course
    - Take a Bayesian methods course