Math 5305 Notes

Introduction

Jesse Crawford

Department of Mathematics
Tarleton State University

Course Topics

- Experimental Design
  - Observational studies vs. experiments
  - Randomization and blinding
  - Confounding variables
- Multiple linear regression model
  - $$Y_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i, \text{ for } i = 1, \ldots, n.$$
  - What are the underlying assumptions of this model?
  - How can we test these assumptions?
  - What goes wrong if the assumptions are violated?
  - If the assumptions are valid, how can we estimate the model parameters and perform hypothesis tests?
  - Variable selection and model building.

- Logistic regression model
  - Output is dichotomous ($Y_i = 0$ or 1).
  - $$g_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$
    $$P[Y_i = 1] = \frac{e^{g_i}}{1 + e^{g_i}}, \text{ for } i = 1, \ldots, n.$$
- Other Multivariate Analysis Techniques
  - Principle components
  - Canonical correlations
  - Factor analysis
  - Discriminant analysis
  - Cluster analysis

- Skills used
  - Critical thinking and reading
  - Formal mathematics (rigorous proofs)
  - Programming (in R and SAS)

Outline

Random Variables

Definition (Informal)

- A *random variable* is a real number whose value is determined randomly.
- Random variables are usually denoted by capital letters, X, Y, U, V, etc.

Definition

The *support* of a random variable is the set of all possible values of that random variable.

Discrete Random Variables

Definition

A random variable is called *discrete* if its support is countable (finite or countably infinite).

Example

- A football player attempts 10 field goals.
- Let X be the number of successful attempts.
- What is the support for X?
- Is X a discrete random variable?

Example

- Let X be the number of phone calls received by a company in one hour.
- What is the support for X?
- Is X a discrete random variable?

Probability Mass Functions

Definition

- Suppose X is a discrete random variable.
- The probability mass function for X is given by

$$f(x) = P[X = x],$$

  for each value of x in the support of X.

Example

- A football player attempts 10 field goals.
- The attempts are statistically independent, and
- The probability of success on each attempt is 0.7.
- Find the p.m.f. for X.
- Find the probability that the player makes exactly 6 field goals.

The Binomial Distribution

Definition

- Let n be a positive integer, and let $p \in [0, 1]$.
- The *binomial distribution* with *parameters* n and p is given by the p.m.f.

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \; x = 0, 1, \ldots, n.$$

Parameters are constants related to a probability distribution.

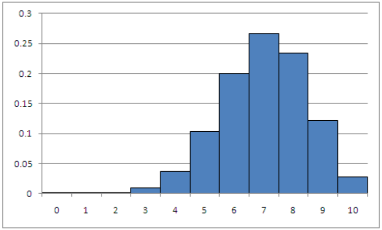


Figure: Binomial distribution with $n = 10$ and $p = 0.7$.

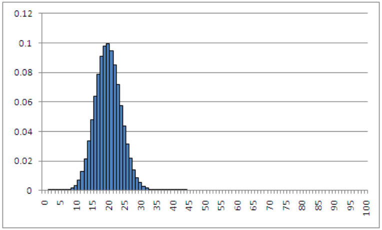


Figure: Binomial distribution with $n = 100$ and $p = 0.2$.

Expected Value, Variance, and Standard Deviation

Definition

- Let X be a random variable with p.m.f. f.
- The *expected value* or *mean* of X is given by

$$E[X] = \mu_X = \sum_{x \in \mathbb{R}} xf(x).$$

- The expected value is the "center of mass" of the distribution, and it tells you the *average value* of the random variable.
- The *variance* of X is

$$\text{Var}[X] = \sigma_X^2 = E[(X - \mu_X)^2] = E[X^2] - E[X]^2.$$

- The *standard deviation* of X is the square root of the variance,

$$\sigma_X = \sqrt{\text{Var}[X]}.$$

- The variance and standard deviation are measures of variation in X.
- The standard deviation provides a rough measure of the spread in the distribution of X.
- It is roughly the average distance from X to its mean.

EV and Variance for Binomial Distributions

- Suppose X has a binomial distribution with parameters n and p.
- Then

$$E[X] = np, \text{ and}$$

$$\mathrm{Var}[X] = np(1-p).$$

Continuous Random Variables

Definition

- Let X be a random variable, and suppose
- $f : \mathbb{R} \to [0, \infty)$, such that

$$P[a < X < b] = \int_a^b f(x)\, dx,$$

for any $a, b \in \mathbb{R}$, such that $a < b$.

- Then X is called a *continuous random variable*, and
- f is its *probability density function*.

The Normal Distribution

Definition

- Suppose $\mu \in \mathbb{R}$, and $\sigma > 0$.
- The *normal distribution* with mean μ and standard deviation σ, is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad -\infty < x < \infty.$$

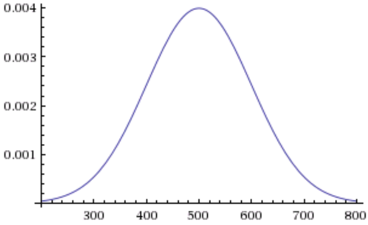


Figure: Normal distribution with $\mu = 500$ and $\sigma = 100$.

Example

- Suppose $X \sim N(500, 100^2)$.
- Find $P[400 < X < 600]$.
- Find $E[X]$ and σ_X.

Proposition

- Let X be a continuous random variable with p.d.f. f.
- Then
$$E[X] = \int_{-\infty}^{\infty} xf(x)\, dx.$$
-
$$E[X^2] = \int_{-\infty}^{\infty} x^2 f(x)\, dx.$$
-
$$\mathrm{Var}[X] = E[X^2] - E[X]^2$$

Standard Normal Distribution

Definition

- The normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$
- is called the *standard normal distribution*.
- Standard normal random variables are usually denoted by Z.

Definition

- Let Z be a standard normal random variable, and
- let $\alpha \in (0, 1)$.
- We define z_α to be the unique number such that

$$P[Z > z_\alpha] = \alpha.$$

α	$z_{\alpha/2}$
0.05	1.96
0.01	2.575

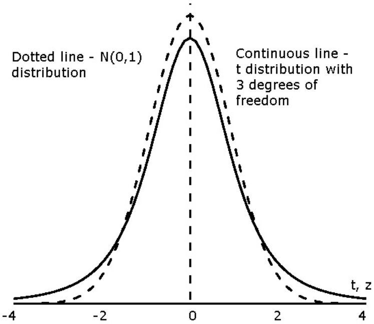
The t-distribution

Definition

- Let r be a positive integer.
- The t-distribution with r degrees of freedom is given by

$$f(t) = \frac{\Gamma((r+1)/2)}{\sqrt{\pi r}\,\Gamma(r/2)} \frac{1}{(1+t^2/r)^{(r+1)/2}}, \quad -\infty < t < \infty.$$

- The t-distribution resembles the $N(0,1)$ distribution, but with fatter tails, and
- the larger the degrees of freedom is, the closer the resemblance is.


Dotted line - N(0,1)
distribution
Continuous line -
t distribution with
3 degrees of
freedom
t, z
-4
-2
0
2
4

Definition

- Let T have a t-distribution with r degrees of freedom.
- let $\alpha \in (0, 1)$.
- We define $t_\alpha(r)$ to be the unique number such that

$$P[T > t_\alpha(r)] = \alpha.$$

α	$z_{\alpha/2}$
0.10	1.645
0.05	1.96
0.01	2.575

α	$t_{\alpha/2}(30)$
0.10	1.697
0.05	2.042
0.01	2.750

Outline

1. Probability

2. Statistics

3. Statistics, by Freedman, Pisani, and Purves

Point Estimation

Example

- Suppose a radioactive sample emits particles, and
- the waiting times between the emissions are
- exponentially distributed with unknown mean θ.
- Let $X_1, \ldots, X_n$ be an independent random sample of waiting times.
- Find the best estimate for θ based on $X_1, \ldots, X_n$.

Important Components of a Statistical Model

- A *population distribution* $f(x; \theta)$.
- The *unknown parameter* θ.
  - Parameters are numbers related to the population.
  - They are constants (not random).
- A *random sample* $X_1, \ldots, X_n$.
  - The X_i's are independent random variables.
  - The distribution of each X_i is given by $f(x; \theta)$.

Maximum Likelihood Estimation

Definition

- The *likelihood function* for a statistical model with population distribution $f(x;\theta)$ is

$$L(\theta, x_1, \ldots, x_n) = f(x_1;\theta) \cdots f(x_n;\theta).$$

- The *maximum likelihood estimator* (MLE) for θ based on the sample $X_1, \ldots, X_n$ is the value of θ that maximizes $L(\theta, X_1, \ldots, X_n)$.
- The MLE is usually denoted by $\hat{\theta}$.
- The MLE is a function of the sample.
- The MLE is a random variable.

Point Estimation for the Normal Distribution

- Consider a random sample $X_1, \ldots, X_n$ from a $N(\mu, \sigma^2)$ population.
- The MLEs for μ and σ^2 are

$$\hat{\mu} = \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i, \text{ and}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

- Because $\hat{\sigma}^2$ is biased, the following estimator is preferred,

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

- In other words, the population mean and variance are estimated by the sample mean and variance.

Test Statistics for the Normal Distribution

Proposition

- Consider a random sample $X_1, \ldots, X_n$ from a $N(\mu, \sigma^2)$ population.
- $$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$
- $$T = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t(n - 1).$$

The Central Limit Theorem

Theorem (5.6-1)

- *Suppose* $X_1, X_2, \ldots$ *is a sequence of IID random variables,*
- *from a distribution with finite mean* μ
- *and finite positive variance* σ^2.
- *Let* $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, *for* $n = 1, 2, \ldots$
- *Then, as* $n \to \infty$,

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \Rightarrow N(0, 1).$$

Informal Statement of CLT

Informal CLT

- Suppose $X_1, \ldots, X_n$ is a random sample
- from a distribution with finite mean μ
- and finite positive variance σ^2.
- Then, if n is sufficiently large,

$$\overline{X} \approx N(\mu, \sigma^2/n), \text{ and}$$

$$\sum_{i=1}^{n} X_i \approx N(n\mu, n\sigma^2).$$

- Conventionally, values of $n \geq 30$ are usually considered sufficiently large, although this text applies the approximation for lower values of n, such as $n \geq 20$.

Finite Population Correction Factor

- Suppose $X_1, \ldots, X_n$ is a random sample
- from a **finite** population with finite mean μ
- and finite positive variance σ^2.
- Assume the population size is N.
- Then, if n is sufficiently large,

$$\overline{X} \approx N\left(\mu, \frac{\sigma^2}{n}\frac{N-n}{N-1}\right).$$

Confidence Intervals

- Let $\alpha \in (0, 1)$ (for example $\alpha = 0.05$).
- Then a $1 - \alpha$ confidence interval for μ is

$$\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right).$$

- This random interval will contain the unknown mean μ with probability $1 - \alpha$.
- If $\alpha = 0.05$, this is a 95% confidence interval, and the probability it contains μ is 95%.
- Alternative way of writing the confidence interval:

$$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

- A more useful confidence interval is

$$\overline{X} \pm t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}.$$

Hypothesis Testing

Example

- Suppose Math SAT scores at a certain university are normally distributed with unknown mean μ and unknown variance σ^2.
- Consider the *hypothesis testing problem*

$$\mathrm{H}_0 : \mu = 500 \text{ vs. } \mathrm{H} : \mu \neq 500.$$

- How can we address this problem using a random sample $X_1, \ldots, X_n$ of n students' Math SAT scores?

- Type I error: Rejecting H_0 when it is true.
- Type II error: Not rejecting H_0 when it is false.
- Can't control the probabilities of both types of errors.
- Instead, we choose $\alpha \in (0, 1)$, called the significance level,
- and require $P[\text{Type I error}] \leq \alpha$.

Hypothesis Testing for the Normal Distribution

- Suppose $X_1, \ldots, X_n$ is a random sample from a $N(\mu, \sigma^2)$ population.
- Let $\mu_0 \in \mathbb{R}$, and consider the testing problem

$$\mathrm{H}_0 : \mu = \mu_0 \text{ vs. } \mathrm{H} : \mu \neq \mu_0.$$

- Testing procedure: reject H_0 if $|Z| \geq z_{\alpha/2}$, where

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}.$$

- A more useful procedure is to reject H_0 if $|T| \geq t_{\alpha/2}(n-1)$, where

$$T = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}.$$

p-values

Once a sample has been collected and a test statistic has been calculated, the p-value of the test can also be calculated.

Definition

The p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming the null hypothesis is true.

This allows for a simple test procedure: reject H_0 if the p-value is less than α.

Related Reading

- *Probability and Statistical Inference*, 8^{th} ed., by Hogg and Tanis.
- My Math 311 and Math 411 notes cover these concepts in much more detail.
- *Introduction to Mathematical Statistics*, by Hogg, McKean, and Craig, for a more rigorous treatment of the same concepts.
- *Probability and Measure*, by Billingsley, for an excellent measure-theoretic treatment of probability.

Outline

Confounding Variables

Definition

- Suppose you are investigating the relationship between the variables X and Y.
- A *confounding variable* is a third variable Z that is related to both X and Y, creating the illusion of a causal relationship between X and Y when there isn't one.

Example

- Men who drink alcohol have higher lung cancer rates.
- Is this strong evidence that alcohol causes cancer?

- “Post hoc ergo propter hoc” fallacy
- “After this, therefore because of this”

Example

- Stimulus package in 2009.
- What was the effect on unemployment?

Randomized Controlled Experiments

- When studying the effect of a treatment, it is necessary to compare a *treatment group*, who receives the treatment, to a *control group*, who does not.
- Subjects should be divided between the treatment group and control group randomly.
- Blinding should be used when appropriate.

- Let p_1 and p_2 be two population proportions, and consider

$$\mathrm{H}_0 : p_1 = p_2 \text{ vs. } \mathrm{H}_1 : p_1 \neq p_2.$$

- Let $\hat{p}_1 = Y_1/n_1$ and $\hat{p}_2 = Y_2/n_2$ be corresponding sample proportions based on **independent** samples of sizes n_1 and n_2, respectively.
- Also, assume that both $n_i\hat{p}_i \geq 5$ and $n_i(1 - \hat{p}_i) \geq 5$, for $i = 1, 2$.
- Decision rule:

$$\text{Reject } \mathrm{H}_0 \text{ if } |Z| \geq z_{\alpha/2}, \text{ where}$$

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \text{ and}$$

$$\hat{p} = \frac{Y_1 + Y_2}{n_1 + n_2}.$$

Observational Studies

Definition

- Controlled Experiment: a study where the investigator assigns subjects to treatment and control groups.
- Observational Study: a study where the investigator does not interact with the subjects being studied. The investigator simply analyzes existing data.

Example

- Smokers (treatment group): higher rates of lung cancer
- Nonsmokers (control group): lower rates of lung cancer
- Is this a controlled experiment or observation study?

- Observational studies can benefit from the use of *homogenous classes*.
- Example: comparing lung cancer rates for male **smokers** of age 55-59 to lung cancer rates for male **nonsmokers** of age 55-59.

Definition

- *Controlling for a variable* means including that variable in a study so it does not distort the relationship between the primary variables being studied.
- In the above smoking/lung cancer study, we are controlling for gender and age.
- Using homogenous classes is one way to control for variables.
- Another method is to include those variables in a statistical model.

Pitfalls of Uncritical Reading

Example

"In a study of clofibrate, 15% of those taking the drug died within the 5 year study, while 25% of those not taking the drug died during the study."

Example

"In a study of Pellagra, the disease was linked to the presence of the blood-sucking fly Simulium."

Example

"In a recent study, it was found that babies exposed to ultrasound in the womb had lower birthweight, on average, than those who were not exposed."

Example

"A study of U.C. Berkeley admissions showed that, over a certain time period, 44% of male applicants were admitted to the graduate school, and only 35% of female applicants were admitted to the graduate school."