**Data Analysis**
Spring Semester, 2023
January 31, 2023
Lecture 3

***Feb 23:*** ***Examination One: Chapters 3, 4, 5, 6, and 7***

Two major points:

In a randomized experiment, a difference in response between the group given an experimental treatment and the group given a control treatment is ***CAUSED*** by the experimental treatment or is a chance event.

In an observational study, correlation shows association, not necessarily causation.

Please follow a news source (New York Times, Wall Street Journal, Newsday, CNN, …) regularly to identify stories and issues that are relevant to this course.

## *Chapter One, Statistics and the Scientific Method:*

The schematic for the scientific method (Figure 1.1): 1. formulate research goal; 2. plan the study (specifically identify variables); 3. collect data; 4. inferences; 5. conclusions (decisions); 6. formulate new goals and return to step 2.

## *Chapter Three, Data Description:*

Your text does not discuss scales of measurement. Wikipedia and your favorite search engine are effective tools to get a fuller definition of scales of measurement. The ***nominal scale*** of measurement is the simplest. The value of a nominal scale variable is actually a verbal characterization. For example, ethnicity or hair color are nominal variables. Traditionally, numbers are used in databases rather than the words. Pie charts, bar charts, and contingency tables are tools used to deal with a nominal scale variable. The ***ordinal scale*** of measurement has ranked values. For example, course grade with values A, B, …, is an example of an ordinal scale variable. Psychology statistics texts often recommend nonparametric statistical procedures. The common statistical procedures that you studied in AMS 310 can be viewed as an approximation to a permutation procedure. The permutation procedure is considered a gold standard statistical procedure in scientific research. In an ***interval scale***, differences have meaning. For example, the net balance in a bank

account (positive for asset and negative for liability) is an example of an interval scale variable. The techniques that you studied in AMS 310 (such as t-tests) can be applied to interval scale variables. A ***ratio scale*** variable is one in which ratios have meaning. For example, the weight of a study participant is a ratio scale variable. Often, a monotonic transformation of a ration scale variable is helpful.

## *Chapter 4, Probability and Probability Distributions*

The theorem that the probability of the union of two events, *A* and *B*, is $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. This leads immediately to Boole's inequality, which states that $P(A \cup B) \leq P(A) + P(B)$ for any two events *A* and *B*.

*Conditional Probability and Bayes' Theorem*

Definition of conditional probability of *A* given *B*: $P(A|B) = \frac{P(A \cap B)}{P(B)}, P(B) > 0.$

The sets *A*, *B*, and *C* are a collection of cover sets if the sample space *S* is such that $S = A \cup B \cup C$, where *A*, *B*, and *C* are disjoint (mutually exclusive or incompatible)

The law of total probability: $P(E) = P(E|A)P(A) + P(E|B)P(B) + P(E|C)P(C).$

Bayes' Theorem: $P(A|E) = \frac{P(E|A)P(A)}{P(E)} = \frac{P(E|A)P(A)}{P(E|A)P(A)+P(E|B)P(B)+P(E|C)P(C)}.$

When the random variable *X* has the binomial distribution on *n* trials with probability of success *p* with $0 < p < 1$, $E(X) = np$, $\text{var}(X) = np(1-p)$, and $\text{var}(X) < E(X)$. When the random variable *Y* has the Poisson distribution with mean $\mu$, $E(Y) = \mu$, $\text{var}(Y) = \mu$, and $E(Y) = \text{var}(Y)$. In lecture, I mentioned "over-dispersed" discrete distributions. For example, if *W* has a negative binomial distribution, $\text{var}(W) > E(W)$.

Example Problem on Bayes' Theorem:

An individual has one of three genotypes called *A*, *B*, and *C*, respectively, for a gene associated with disease *X*. The probability that an individual has genotype *A* is 0.64; the probability that an individual has genotype *B* is 0.32; and the probability that an individual has genotype *C* is 0.04. The probability that an individual with the *A* genotype is affected with disease *X* is 0.05. The probability

that an individual with the *B* genotype is affected with disease *X* is 0.80. The probability that an individual with the *C* genotype is affected with disease *X* is 0.99.

a. What is the probability that an individual is affected with disease *X*?
b. Given that an individual has disease *X*, what is the probability that the individual is genotype *B*?

Normal Probability Distribution.

Let the random variable *X* be normally distributed with expected value $\mu$ and variance $\sigma^2$.
That is, $X \sim N(\mu, \sigma^2)$. Any probability calculation about a normal distribution can be transformed to a calculation with a standard normal:

$$P(X \leq a) = P\left(\frac{X-\mu}{\sigma} \leq \frac{a-\mu}{\sigma}\right) = \Phi\left(\frac{a-\mu}{\sigma}\right),$$

where $\Phi$ is the cumulative distribution function of the standard normal. That is, $\Phi(z) = \Pr\{Z \leq z\}$, where $Z \sim N(0,1)$.

Key Percentiles of the Standard Normal:

- $P(Z \leq -2.576) = \Phi(-2.576) = 0.005$
- $P(Z \leq -2.326) = \Phi(-2.326) = 0.01$
- $P(Z \leq -1.960) = \Phi(-1.960) = 0.025$
- $P(Z \leq -1.645) = \Phi(-1.645) = 0.05$
- $P(Z \leq -1.282) = \Phi(-1.282) = 0.10$
- $P(Z \leq -0.6745) = \Phi(-0.6745) = 0.25$

Definition of Expected Value of a Discrete Random Variable:

$$E(X) = \sum_{APV} xP(X = x),$$

where *APV* means to sum over all possible values of the discrete random variable.
Expectation is a linear operator.

Definition of variance of the random variable *X*: $var(X) = E((X - EX)^2)$

Calculating the variance of the Bernoulli random variable:

$$(0-p)^2(1-p) + (1-p)^2 p = p(1-p)$$

Important identity:

$$var(X) = E((X - EX)^2) = E(X^2) - (EX)^2$$



Abstract problem:看视频

The proof starts with $E[(X-EX)^2] = E[X^2 - 2X(EX) + (EX)^2]$

$$= E[X^2] - E[2X(EX)] + E[(EX)^2]$$
$$= E[X^2] - 2(EX)E[X] + (EX)^2$$
$$= E[X^2] - 2(EX)^2 + (EX)^2$$
$$= E(X^2) - (EX)^2.$$

会考

Example Problem:

The random variables $W_1$ and $W_2$ are a random sample of 2 drawn from the random variable $W$
which has expected value $\mu_W$ and standard deviation $\sigma_W > 0$. Find $E(W_1 - W_2)$ and $E((W_1 - W_2)^2)$.

Solution:

$$E(W_1 - W_2) = E(W_1) - E(W_2) = \mu_W - \mu_W = 0;$$

These steps follow from the linear operator property of expectation.
To find $E((W_1 - W_2)^2)$:

$$E((W_1 - W_2)^2) = E[W_1^2 - 2W_1W_2 - W_2^2]$$
$$= E[W_1^2] - E[2W_1W_2] - E[W_2^2].$$

First, since $var(W) = E[W^2] - [E(W)]^2$, $E[W^2] = var(W) + [E(W)]^2 = \sigma_W^2 + \mu_W^2$.
Second, $E[W_1W_2] = E(W_1)E(W_2)$, since $W_1$ and $W_2$ are independent.
Then, $E[W_1^2] - E[2W_1W_2] - E[W_2^2] = \sigma_W^2 + \mu_W^2 - 2E(W_1)E(W_2) + \sigma_W^2 + \mu_W^2$
Combining, $E((W_1 - W_2)^2) = 2\sigma_W^2 + 2\mu_W^2 - 2\mu_W^2$

$$= 2\sigma_W^2.$$

In a random sample of size 2, $\sigma_W^2 = \frac{E((W_1-W_2)^2)}{2}$. An unbiased estimate of $\sigma_W^2$ is $\frac{(W_1-W_2)^2}{2}$.

不懂

## *Chapter Five*

Distribution of Sample Mean:

- Let $Y_1, Y_2, \ldots, Y_n$ be a random sample of size *n* from *Y* which has the distribution $N(\mu, \sigma^2)$.

  ***THEN*** the distribution of $\bar{Y}_n = \frac{\sum_{i=1}^{n} Y_i}{n}$, the sample mean, is $N(\mu, \sigma^2/n)$.

- Central Limit Theorem (CLT): When $Y_1, Y_2, \ldots, Y_n$ is a random sample of size *n* from *Y* which has expected value $\mu$ and variance $\sigma^2 < \infty$, then the distribution of $\bar{Y}_n = \frac{\sum_{i=1}^{n} Y_i}{n}$ is asymptotically $N(\mu, \sigma^2/n)$. For a random sample of size *n*, it is always true that $E(\bar{Y}_n) = \mu$ and $\text{var}(\bar{Y}_n) = \frac{\sigma^2}{n}$. The CLT allows probability calculations that increase in accuracy as the sample size increases.

Testing a Statistical Hypothesis (Variance Known)

- Null hypothesis: $H_0 : E(Y) = \mu_0$
- Alternative Hypothesis: $H_1 : E(Y) \neq \mu_0$
- Level of significance $\alpha$
- Type I error: reject a null hypothesis that is true.
- The probability of a Type I error is $\alpha$. Formally, $\Pr_0\{\text{Reject } H_0\} = \alpha$.
- Test statistic. Null distribution is the distribution of the test statistic under the null hypothesis.
- Type II error: accept a null hypothesis that is false.
- Typically, $\alpha$ is set to a small number (0.05 or 0.01), and *n* is chosen so that $\beta = \Pr_1\{\text{Accept } H_0\}$, where $\beta$ is dependent on a setting of the alternative hypothesis, is small. Alternative distribution: distribution of the test statistic under the setting of the alternative hypothesis.

Example of Statistical Hypothesis (Variance Known)

- A research team took a sample of 8 observations from the random variable *Y*, which had a normal distribution $N(\mu, \sigma^2 = 625)$. They

observed $\bar{y}_8 = 43.2$, where $\bar{y}_8$ is the average of the eight sampled observations. Test the null hypothesis that $H_0 : E(Y) = 50$ against the alternative $H_1 : E(Y) \neq 50$ at the 0.10, 0.05, and 0.01 levels of significance.

- The test statistic is $\bar{Y}_8$.
- The null distribution is $N(50, 625/8 = 78.125 = 8.84^2)$
- Put $\bar{Y}_8$ in standard score form: $Z = \dfrac{\bar{Y}_8 - \mu_0}{\sigma / \sqrt{8}}$.
- If $|Z| \geq 1.645$, reject $H_0 : E(Y) = \mu_0$ at $\alpha = 0.10$. If $|Z| \geq 1.960$, reject $H_0 : E(Y) = \mu_0$ at $\alpha = 0.05$. If $|Z| \geq 2.576$, reject $H_0 : E(Y) = \mu_0$ at $\alpha = 0.01$.
- Calculate the standard score form of *Z* for the data given in the problem: $z = \dfrac{43.2 - 50}{8.84} = -0.769$
- Since $|z| < 1.645$, accept $H_0 : E(Y) = 50$ at $\alpha = 0.10$. Of course, one should accept $H_0 : E(Y) = 50$ at $\alpha = 0.05$ and $\alpha = 0.01$ as well.

Testing a Statistical Hypothesis (Variance Unknown).

- We cannot put $\bar{Y}_n$ in standard score form: $Z = \dfrac{\bar{Y}_n - \mu_0}{\sigma / \sqrt{n}}$ because we do not know $\sigma$.
- Instead, we use an estimate of $\sigma^2$, $\hat{\sigma}^2 = S^2 = \dfrac{\sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2}{n-1}$, which has $n-1$ degrees of freedom.
- We put $\bar{Y}_n$ in studentized standard score form: $T_{n-1} = \dfrac{\bar{Y}_n - \mu_0}{\hat{\sigma} / \sqrt{n}}$.
- Student showed that the percentiles from the standard normal (here 1.645, 1.960, and 2.576) had to be stretched. The amount of stretching is determined by the number of degrees of freedom in $\hat{\sigma}^2 = S^2 = \dfrac{\sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2}{n-1}$; here, $n-1$ degrees of freedom. These values are tabulated and will be given to you in your examinations.
- Example problem: Chapter 5 Study Guide, Problem 4:

  A research team took a sample of 8 observations from the random variable *Y*, which had a normal distribution $N(\mu, \sigma^2)$. They observed $\bar{y}_8 = 43.2$, where

$\bar{y}_8$ is the average of the eight sampled observations and $s^2 = 517.5$ is the observed value of the unbiased estimate of $\sigma^2$, based on the sample values. Test the null hypothesis that $H_0 : E(Y) = 50$ against the alternative $H_1 : E(Y) \neq 50$ at the 0.10, 0.05, and 0.01 levels of significance.

- The degrees of freedom is $n - 1 = 8 - 1 = 7$.
- The studentized test statistic is $t_7 = \frac{43.2 - 50}{\sqrt{(517.5/8}} = \frac{-6.8}{\sqrt{64.7}} = \frac{-6.8}{8.04} = -0.845$
- Find the student t stretches for 1.645, 1.960, 2.576. They are 1.895, 2.365, 3.499.
- Make your decision. Here, it is to accept at the 0.10 level of significance (and of course at 0.05 and 0.01 as well) since $|-0.845| < 1.895$.

Confidence Interval, Variance Known

- The formal test of a null hypothesis addresses whether a single value (here a value for the expected value of the sampled random variable) is consistent with the data.
- Most researchers prefer a statement of what the data does show.
- The confidence interval is such a statement.
- The 99% confidence interval for *E(Y)* is $\bar{y}_n \pm 2.576 \frac{\sigma}{\sqrt{n}}$.
- In our first example is: $43.2 \pm 2.576(8.84) = 43.2 \pm 22.8$

Confidence Interval, Variance Unknown

- As always, we use the data $S^2$ to estimate $\sigma^2$ and stretch our normal percentile (here 2.576) using the degrees of freedom of $S^2$. The stretch of 2.576 is 3.499.
- The 99% confidence interval for *E(Y)* is $\bar{y}_n \pm t_{n-1,2.576} \frac{\hat{\sigma}}{\sqrt{n}}$.
- In the example problem, the 99% confidence interval for *E(Y)* is $\bar{y}_8 \pm t_{7,2.576} \frac{\hat{\sigma}}{\sqrt{n}} = 43.2 \pm 3.499 \sqrt{\frac{517.5}{8}} = 43.2 \pm 3.499(8.04) = 43.2 \pm 28.1$.

*Paired t-test*

The most common application of the paired t-test is a comparison of the post-training score of a participant in a study with the same participant's pre-

training score. The idea of the paired t-test is to calculate the difference (here post score-pre score) for each participant. This data is used in the one-sample t-test of Chapter 5 to test the null hypothesis that the expected post training score is equal to the expected pre training score.

*Chapter 6 Study Guide, Problem 5*

A research time wished to estimate the reduction of the density of contaminant in a liquid due to filtering the liquid. They filtered four samples, called A, B, C, and D. Find the 99% confidence interval for the expected reduction in the density of contaminant using the data in the table below:

| Sample | Density of Contaminant before Filtering | Density of Contaminant after Filtering | Difference | Deviation of Difference | Deviation Squared |
|---|---|---|---|---|---|
| A | 132 | 87 | 45 | -4.75 | 22.5625 |
| B | 205 | 163 | 42 | -7.75 | 60.0625 |
| C | 81 | 35 | 46 | -3.75 | 14.0625 |
| D | 423 | 357 | 66 | 16.25 | 264.0625 |

Solution: The four differences of *D*=before filtering – after filtering are: 132-87=45, 42, 46, and 66. Then $\bar{d}_4 = \frac{45+42+46+66}{4} = 49.75$ and $s_D^2 = 120.25$ on 3 degrees of freedom. The 99% confidence interval for the expected reduction in density is from 17.7 to 81.8.