NC STATE UNIVERSITY

CSC-591: Foundations of Data Science
T/Th. 12:50-2:05pm. EBI-1005.

Ranga Raju Vatsavai
Chancellors Faculty Excellence Associate Professor in Geospatial Analytics
Department of Computer Science, North Carolina State University (NCSU)
Associate Director, Center for Geospatial Analytics, NCSU
&
Joint Faculty, Oak Ridge National Laboratory (ORNL)

W6: 9/29/15-10/1/15

NC STATE UNIVERSITY

Administrative

- Updated Weekly Schedule (on Moodle)
- 1st Midterm: 10/6/15
- Additional Reading Materials
  - Logistic Regression, by Kleinbaum (Springer, Through NCSU Library)
  - Computing Primer for Applied Linear Regression, 4ed, Using R. http://z.umn.edu/alrprimer

NC STATE UNIVERSITY

Today

- Midterm-1 Review

NC STATE UNIVERSITY

Basic Rules

- Exam: 1 Hour (10 minute bonus)
- Every one should be seated by 12.50pm
- Paper distribution starts at 12.50pm
- Exam starts at: 12.55pm (don't write anything before that)
- Exam ends at: 2.05pm (should close your paper, turn it upside down)
- Please note, if you were caught writing after 2.05pm, you will loose 10% of exam points.
- Don't look at your neighbors answers (if caught, both will be asked to leave exam hall instantly, may lead to other penalties as well).
- Don't ask questions after exam starts, if you are in doubt, simply make your best judgment (write it down clearly), in such cases, instructor/TA decision is final.
- Exam supervisor: TA; there will be couple of more co-supervisors

NC STATE UNIVERSITY

What's Allowed and Not Allowed

- 1-page of written/typed notes (standard US A4 size paper; you can use both sides)
- Regular calculator
- No other electronic (including cell phone as calculator) or paper products are allowed
- Don't put any other personal items on the table
- Statistical tables will be supplied with exam paper, so don't bring your own stuff

NC STATE UNIVERSITY

General Question(s)

- Will test your basic understanding of various data science topics from first couple of lectures
  - Your answers should be straight forward
- Examples
  - What is the impact of poor-data quality on a classification/prediction model?
  - Define attribute types and give example of each

NC STATE UNIVERSITY

Probability

- Basic set operations
  - Complement, union, intersection, difference, subset, disjoint, partition
- Sample space, events, …
- Definition of probability distribution, P, and three axioms
- Counting

NC STATE UNIVERSITY

Probability

- Independent events, important, why?
  - Multiplication
  - If we flip coin twice, probability of two heads = ½ x ½ = ¼
- Conditional probability, P(A|B) = P(AB)/P(B)
  - Think P(A|B) as the fraction of times A occurs among those in which B occurs

NC STATE UNIVERSITY

Bayes Theorem

Theorem 2.16 (Bayes' Theorem.) *Let $A_1, \ldots, A_k$ be a partition of Ω such that $\mathbb{P}(A_i) > 0$ for each i. If $\mathbb{P}(B) > 0$ then, for each $i = 1, \ldots, k$,*

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_j \mathbb{P}(B|A_j)\mathbb{P}(A_j)}. \quad (2.5)$$

- Given:
  - A doctor knows that meningitis causes stiff neck 50% of the time
  - Prior probability of any patient having meningitis is 1/50,000
  - Prior probability of any patient having stiff neck is 1/20
- If a patient has stiff neck, what's the probability he/she has meningitis?

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

NC STATE UNIVERSITY

Probability

- Radom Variables
  - Discrete
  - Continuous

NC STATE UNIVERSITY

Statistical Distributions

- Discrete
  - Probability mass function (pmf)
- Examples
  - Bernoulli distribution is the probability distribution of a random variable which takes value 1 with success probability p and value 0 with failure probability q=1-p.

$$f(k;p) = p^k(1-p)^{1-k} \quad \text{for } k \in \{0,1\}.$$

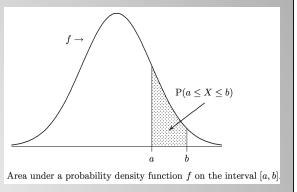
NC STATE UNIVERSITY

Statistical Distributions

- Continuous
  - Probability density function (pdf)
- Examples
  - Continuous uniform distribution $f(x) = \frac{1}{\beta-\alpha} \quad \text{for } \alpha \le x \le \beta.$
  - Normal distribution

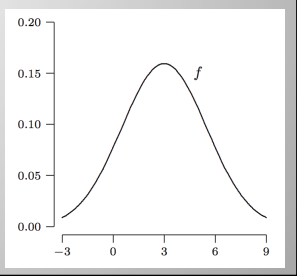
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

    - 68%(μ±σ), 95%(μ±2σ), 99.7%(μ ±3σ)



Area under a probability density function f on the interval $[a, b]$

$$\mathrm{P}(a \le X \le b) = \int_a^b f(x)\,\mathrm{d}x.$$



NC STATE UNIVERSITY

Expected Values

- The expected value is also known as the expectation, mathematical expectation, EV, mean, or first moment.

Outcomes	a_1	a_2	...	a_n
Probability	p_1	p_2		p_n

- $E[X] = a_1p_1 + a_2p_2 + \ldots + a_np_n = \Sigma_i a_i p(a_i)$

DEFINITION. The *expectation* of a discrete random variable X taking the values $a_1, a_2, \ldots$ and with probability mass function p is the number

$$\mathrm{E}[X] = \sum_i a_i \mathrm{P}(X = a_i) = \sum_i a_i p(a_i).$$

NC STATE UNIVERSITY

Expected Values

- EV of continuous r.v.

DEFINITION. The *expectation* of a continuous random variable X with probability density function f is the number

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x.$$

NC STATE UNIVERSITY

Variance

DEFINITION. The *variance* $\mathrm{Var}(X)$ of a random variable X is the number

$$\mathrm{Var}(X) = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right].$$

AN ALTERNATIVE EXPRESSION FOR THE VARIANCE. For any random variable X,

$$\mathrm{Var}(X) = \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2.$$

NC STATE UNIVERSITY

Example of E.V.

- The random variable is given by the following PDF.

$$f(x) = \begin{cases} 2(1-x) & \text{if } 0 \le x \le 1, \\ 0 & \text{otherwise} \end{cases}$$

1. First verify that f(x) is valid PDF.

   2(1-x) = 2 - 2x ≥ 0 precisely when x ≤ 1; thus f(x) is everywhere nonnegative

NC STATE UNIVERSITY

Example

1. Check if f(x) has unit area under its graph

$$\int_{-\infty}^{\infty} f(x)dx = 2\int_{0}^{1}(1-x)dx = 2\left(x - \frac{x^2}{2}\right)\Big|_0^1 = 1$$

2. Therefore, f(x) is a valid PDF.
3. Now compute Expected value

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x.$$

NC STATE UNIVERSITY

Example

1. Mean

$$\begin{aligned}\mathbb{E}(X) &= \int_{-\infty}^{\infty} x f(x)dx \\ &= \int_{0}^{1} x\Big[2(1-x)\Big]dx \\ &= 2\int_{0}^{1}\left(x - x^2\right)dx \\ &= 2\left(\frac{x^2}{2} - \frac{x^3}{3}\right)\Big|_0^1 \\ &= 1/3\end{aligned}$$

2. Compute Variance $\mathrm{Var}(X) = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right]$

NC STATE UNIVERSITY

Example

- Variance

$$
\begin{aligned}
\mathrm{Var}(X) &= \int_{-\infty}^{\infty} \big(x - \mathbb{E}(X)\big)^2 f(x)dx \\
&= \int_0^1 \big(x - 1/3\big)^2 \cdot 2(1-x)dx \\
&= 2\int_0^1 \big(x^2 - \frac{2}{3}x + \frac{1}{9}\big)(1-x)dx \\
&= 2\int_0^1 \big(-x^3 + \frac{5}{3}x^2 - \frac{7}{9}x + \frac{1}{9}\big)dx \\
&= 2\big(-\frac{1}{4}x^4 + \frac{5}{9}x^3 - \frac{7}{18}x^2 + \frac{1}{9}x\big)\Big|_0^1 \\
&= 2\big(-\frac{1}{4} + \frac{5}{9} - \frac{7}{18} + \frac{1}{9}\big) \\
&= \frac{1}{18}
\end{aligned}
$$

NC STATE UNIVERSITY

Parameter Estimation

- Maximum Likelihood Estimation

NC STATE UNIVERSITY

Maximum Likelihood Estimation

- General outline for single parameter
  - Write down the likelihood function: L(θ)
  - Maximize likelihood (difficult due to product)
    - Log-likelihood (monotonic)
    - Take ln (natural log)
    - Differentiate l(θ) with respect to the parameter (θ)
    - Set derivative 0 and solve resulting equation
    - Check this is maximum (by taking 2nd derivative) (generally we don't need, e.g., uni-modal Gaussian)

NC STATE UNIVERSITY

Likelihood Function

Let $X_1, \ldots, X_n$ have joint pmf or pdf

$$f(x_1, x_2, \ldots, x_n; \theta_1, \ldots, \theta_m) \tag{7.6}$$

where the parameters $\theta_1, \ldots, \theta_m$ have unknown values. When $x_1, \ldots, x_n$ are the observed sample values and (7.6) is regarded as a function of $\theta_1, \ldots, \theta_m$, it is called the **likelihood function**. The maximum likelihood estimates $\hat{\theta}_1, \ldots, \hat{\theta}_m$ are those values of the θ_i's that maximize the likelihood function, so that

$$f(x_1, x_2, \ldots, x_n; \hat{\theta}_1, \ldots, \hat{\theta}_m) \geq f(x_1, x_2, \ldots, x_n; \theta_1, \ldots, \theta_m) \text{ for all } \theta_1, \ldots, \theta_m$$

When the X_i's are substituted in place of the x_i's, the **maximum likelihood estimators** (mle's) result.

NC STATE UNIVERSITY

MLE Examples

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

For $X_1, X_2, \ldots, X_n$ iid Poisson random variables will have a joint frequency function that is a product of the marginal frequency functions, the log likelihood will thus be:

$$\begin{aligned} l(\lambda) &= \sum_{i=1}^{n}(X_i log\lambda - \lambda - logX_i!) \\ &= log\lambda \sum_{i=1}^{n} X_i - n\lambda - \sum_{i=1}^{n} logX_i! \end{aligned}$$

We need to find the maximum by finding the derivative:

$$l'(\lambda) = \frac{1}{\lambda}\sum_{i=1}^{n} x_i - n = 0$$

Suppose $X_1, \ldots, X_n$ is a random sample from an exponential distribution with parameter λ. Because of independence, the likelihood function is a product of the individual pdf's:

$$f(x_1, \ldots, x_n; \lambda) = (\lambda e^{-\lambda x_1}) \cdot \cdots \cdot (\lambda e^{-\lambda x_n}) = \lambda^n e^{-\lambda \Sigma x_i}$$

The ln(likelihood) is

$$\ln[f(x_1, \ldots, x_n; \lambda)] = n\ln(\lambda) - \lambda \sum x_i$$

Equating $(d/d\lambda)$[ln(likelihood)] to zero results in $n/\lambda - \Sigma x_i = 0$, or $\lambda = n/\Sigma x_i = 1/\bar{x}$. Thus the mle is $\hat{\lambda} = 1/\bar{X}$;

NC STATE UNIVERSITY

MLE Examples

If $X_1, X_2, \ldots, X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$ random variables their density is written:

$$f(x_1, \ldots, x_n|\mu, \sigma) = \prod_{i}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}[\frac{x_i - \mu}{\sigma}]^2\right)$$

Regarded as a function of the two parameters, μ and σ this is the likelihood:

$$\ell(\mu, \sigma) = -nlog\sigma - \frac{n}{2}log2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)$$

$$\frac{\partial \ell}{\partial \sigma} = -\frac{n}{\sigma} + \sigma^{-3}\sum_{i=1}^{n}(x_i - \mu)^2$$

so setting these to zero gives $\bar{X}$ as the mle for μ, and $\hat{\sigma}^2$ as the usual.

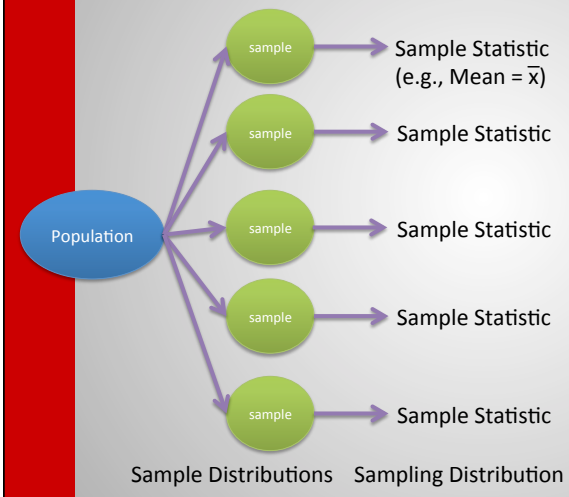
NC STATE UNIVERSITY

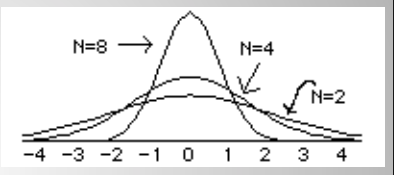
Key properties of MLE

- MLE has nice asymptotic properties
  - Consistent
  - Asymptotically normal
  - Efficiency

NC STATE UNIVERSITY

Sampling Distribution and CLT





The sd of the sampling distribution of the mean is called the standard error of the mean.

$$\sigma_m = \frac{\sigma}{\sqrt{N}}$$

CLT: Given a population with μ and σ^2, the sampling distribution of the mean approaches a normal distribution $\bar{x}$ ~ N(mean = μ, and SE = s/√n)

NC STATE UNIVERSITY

Example

- Let X be a random variable with μ = 10 and σ = 4. A sample of size 100 is taken from this population. Find the probability that the sample mean of these 100 observations is less than 9.
  - We can apply CLT : states that x̄ follows approximately normal: $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$
  - To find probability (area under standard normal), first apply z transformation: $z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$

NC STATE UNIVERSITY

Example

- We have:
  - μ = 10 and σ = 4, n = 100 $z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$
- P(x̄ < 9) = P(z < (9-100) ÷ (4/√100)) = P(z < -2.5)
- Check value from standard normal probabilities table
  - P(z < -2.5) = 0.0062

NC STATE UNIVERSITY

Confidence Intervals

- For quantifying the quality of a particular estimate, we can compute a *confidence interval* (*CI*) around the estimate $\hat{\theta}_n$, such that this interval covers the true value θ with some high probability 1 − α. The narrower this interval, the closer we are to the true value, with high probability.
- CI of population mean: Computed as the sample mean ± a margin of error
  - This is the critical value corresponding to the middle XY% of the normal distribution times the standard error of the sampling distribution

$$\bar{x} \pm z^* (s)/\sqrt{n}$$

NC STATE UNIVERSITY

General Outline for Finding CI

- (1) Identify sample statistic (e.g., mean)
- (2) Select a confidence level (e.g., 90%, 95%, …)
- (3) Compute margin of error:
  - Critical value * Standard deviation of statistic

  OR
  - Critical value * Standard error of statistic
- (3.1) Computing Critical Value:
  - The central limit theorem states that the sampling distribution of a statistic will be nearly normal
  - When the sampling distribution is nearly normal, the critical value can be expressed as a z score.

NC STATE UNIVERSITY

General Outline for Finding CI

- (3.1) Computing Critical Value (continued):
  - When the sampling distribution is nearly normal, the critical value can be expressed as a z score.
    - First compute following quantities:
      - $\alpha = 1 - (\text{confidence level} / 100)$
      - critical probability (p^*): $p^* = 1 - \alpha/2$
    - To express critical value as z score: find the z score having a cumulative probability equal to the critical probability (p^*)
- (4) Finally, CI = Sample statistic ± Margin of error

NC STATE UNIVERSITY

Example

- A sample of 30 students is drawn from CSC-522 population of 100. Average weight of sample is 150 pounds and standard deviation is 20 pounds. Compute the 95% CI.
- Use the 4-step solution outlined in previous slides to find the answer

NC STATE UNIVERSITY

Hypothesis Testing

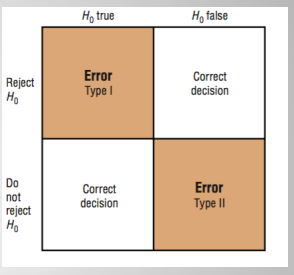
Two-tailed test	Right-tailed test	Left-tailed test
$H_0: \mu = k$ $H_1: \mu \neq k$	$H_0: \mu = k$ $H_1: \mu > k$	$H_0: \mu = k$ $H_1: \mu < k$

- Null hypothesis is always stated with the equals sign
- When a researcher conducts a study, he or she is generally looking for evidence to support a claim. Therefore, the claim should be stated as the alternative hypothesis, i.e., using < or > or ≠. Because of this, the alternative hypothesis is sometimes called the research hypothesis.

NC STATE UNIVERSITY

Possible Outcomes of a Hypothesis Test

- There 4 possible outcomes
- A type I error occurs if you reject the null hypothesis when it is true.
- A type II error occurs if you do not reject the null hypothesis when it is false.



NC STATE UNIVERSITY

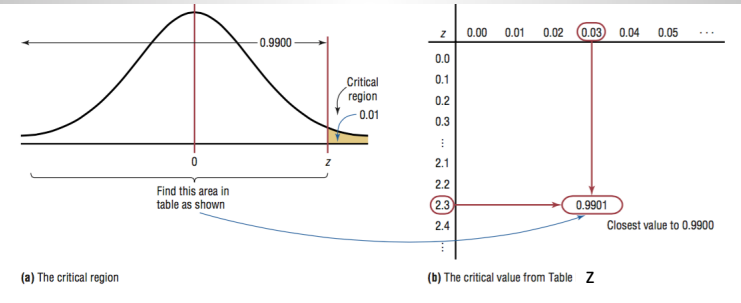
Level of Significance

- The level of significance is the maximum probability of committing a type I error. This probability is symbolized by α. That is, P(type I error) = α.
- *P*(type II error) = β.
  - In most hypothesis-testing situations, β cannot be easily computed; however, α and β are related in that decreasing one increases the other.
- Statisticians generally agree on using three arbitrary significance levels: the 0.10, 0.05, and 0.01 levels.
  - That is, if the null hypothesis is rejected, the probability of a type I error will be 10%, 5%, or 1%, depending on which level of significance is used.

NC STATE UNIVERSITY

Critical Values

- The critical or rejection region is the range of values of the test value that indicates that there is a significant difference and that the null hypothesis should be rejected.
- The noncritical or nonrejection region is the range of values of the test value that indicates that the difference was probably due to chance and that the null hypothesis should not be rejected.
- Finding C.V. for α = 0.01 (right-tailed test); Answer = 2.3+0.03



NC STATE UNIVERSITY

Hypothesis Testing Procedure

Step 1	State the hypotheses and identify the claim.
Step 2	Find the critical value(s) from the appropriate table
Step 3	Compute the test value.
Step 4	Make the decision to reject or not reject the null hypothesis.
Step 5	Summarize the results.

NC STATE UNIVERSITY

z Test

The ***z* test** is a statistical test for the mean of a population. It can be used when $n \geq 30$, or when the population is normally distributed and σ is known.

The formula for the z test is

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

where

$\overline{X}$ = sample mean
μ = hypothesized population mean
σ = population standard deviation
n = sample size

- General form of hypothesis test

$$\frac{\text{(observed value – expected value)}}{\text{standard error}}$$

NC STATE UNIVERSITY

Professors Salaries

- A researcher reports that the average salary of assistant professors is more than $42,000. A sample of 30 assistant professors has a mean salary of $43,260. For α = 0.05, test the claim that assistant professors earn more than $42,000 per year. The standard deviation of the population is $5230.

NC STATE UNIVERSITY

Solution

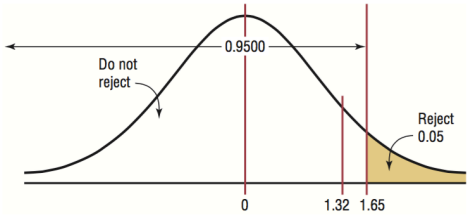
Step 1 State the hypotheses and identify the claim.
H_0: μ = $42,000 and H_1: μ > $42,000 (claim)

Step 2 Find the critical value. Since $\alpha = 0.05$ and the test is a right-tailed test, the critical value is $z = +1.65$.

Step 3 Compute the test value.

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\$43{,}260 - \$42{,}000}{\$5230/\sqrt{30}} = 1.32$$

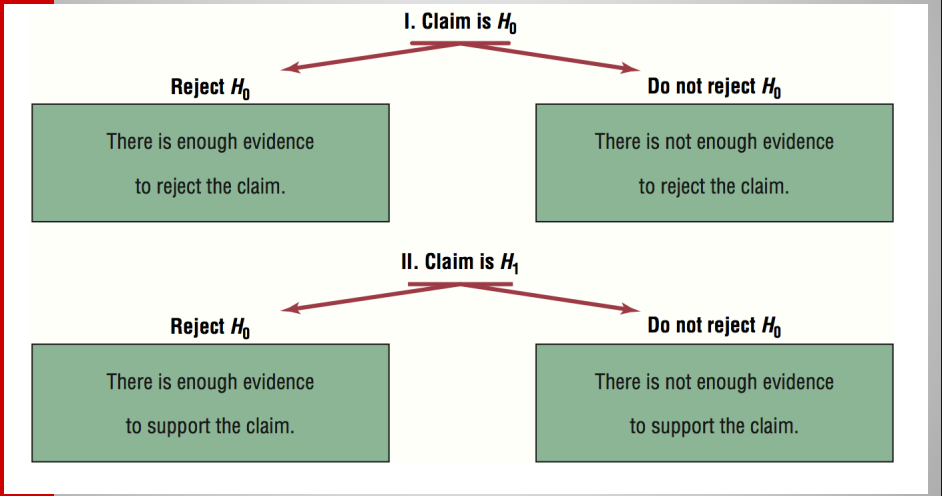
Step 4 Make the decision. Since the test value, +1.32, is less than the critical value, +1.65, and is not in the critical region, the decision is to not reject the null hypothesis. This test is summarized in Figure .



Step 5 Summarize the results. There is not enough evidence to support the claim that assistant professors earn more on average than $42,000 per year.

NC STATE UNIVERSITY

How to Write Summary Statement



NC STATE UNIVERSITY

t distribution

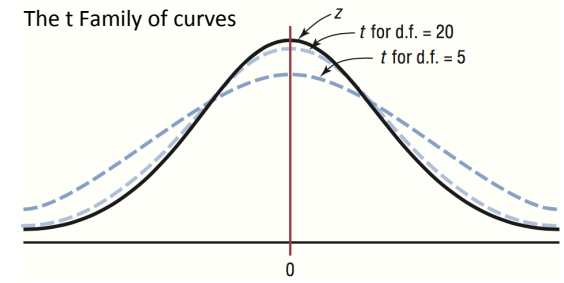
- The *t* distribution is similar to the standard normal distribution in the following ways
  - It is bell-shaped.
  - It is symmetric about the mean.
  - The mean, median, and mode are equal to 0 and are located at the center of the distribution.
  - The curve never touches the *x* axis.
- However, *t* distribution differs from the standard normal distribution in the following ways
  - The variance is greater than 1.
  - The *t* distribution is a family of curves based on the *degrees of freedom*, which is a number related to sample size.
  - As the sample size increases, the *t* distribution approaches the normal distribution.

NC STATE UNIVERSITY

Degrees of freedom

- The degrees of freedom are the number of values that are free to vary after a sample statistic has been computed, and they tell the researcher which specific curve to use when a distribution consists of a family of curves.

For e.g, if the mean of 5 values is 10, then 4 of the 5 values are free to vary. But once 4 values are selected, the fifth value must be a specific number to get a sum of 50, since 50÷5 = 10. Hence, the degrees of freedom are 5-1=4, and this value tells the researcher which *t* curve to use.



NC STATE UNIVERSITY

t Test for a Mean

The ***t* test** is a statistical test for the mean of a population and is used when the population is normally or approximately normally distributed, σ is unknown.

The formula for the *t* test is

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

The degrees of freedom are d.f. $= n - 1$.

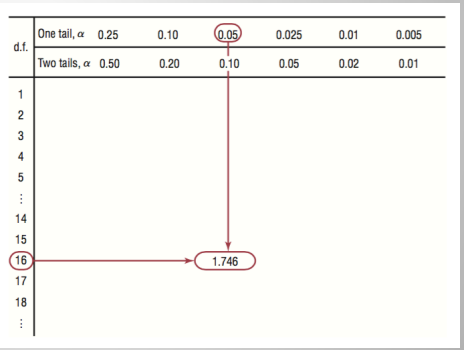
- Like z; the critical values for the *t* test are given in a standard table

NC STATE UNIVERSITY

Example

- Find the critical *t* value for $\alpha = 0.05$ with d.f. = 16 for a right-tailed *t* test.

Find the 0.05 column in the top row and 16 in the left-hand column. Where the row and column meet, the appropriate critical value is found; it is 1.746.



d.f.	One tail, α	0.25	0.10	0.05	0.025	0.01	0.005
	Two tails, α	0.50	0.20	0.10	0.05	0.02	0.01
1							
2							
3							
4							
5							
⋮							
14							
15							
16				1.746			
17							
18							
⋮							

NC STATE UNIVERSITY

Example

- A medical investigation claims that the average number of infections per week at a hospital in southwestern Pennsylvania is 16.3. A random sample of 10 weeks had a mean number of 17.7 infections. The sample standard deviation is 1.8. Is there enough evidence to reject the investigator's claim at $\alpha = 0.05$?

NC STATE UNIVERSITY

Solution

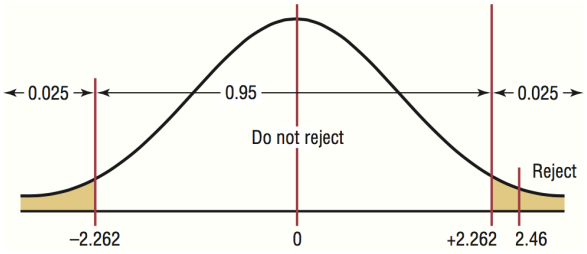
Step 1 H_0: $\mu = 16.3$ (claim) and H_1: $\mu \neq 16.3$.

Step 2 The critical values are $+2.262$ and -2.262 for $\alpha = 0.05$ and d.f. $= 9$.

Step 3 The test value is

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{17.7 - 16.3}{1.8/\sqrt{10}} = 2.46$$

Step 4 Reject the null hypothesis since $2.46 > 2.262$. See Figure



Step 5 There is enough evidence to reject the claim that the average number of infections is 16.3.

10

NC STATE UNIVERSITY

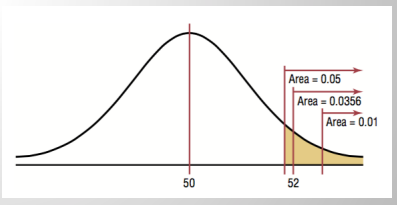
P-Value Method

The ***P*-value** (or probability value) is the probability of getting a sample statistic (such as the mean) or a more extreme sample statistic in the direction of the alternative hypothesis when the null hypothesis is true.

- H_0: μ = 50; H_1: μ >50; Let's say P-value for a statistical test is 0.0356, then probability of getting a sample mean of 52 or grater is 0.0356 if the true population mean is 50 (for given sample size and s.d.

The relationship between the *P*-value and the α value can be explained in this manner.

For *P* = 0.0356, the null hypothesis would be rejected at α=0.05 but not at α=0.01



NOTE: When the hypothesis test is two-tailed, the area in one tail must be doubled. For a two-tailed test, if α=0.05 and the area in one tail is 0.0356, the *P*-value will be 2(0.0356)=0.0712. Now, do you reject H_0?

NC STATE UNIVERSITY

Summary of P-Value Procedure

Step 1	State the hypotheses and identify the claim.
Step 2	Compute the test value.
Step 3	Find the *P*-value.
Step 4	Make the decision.
Step 5	Summarize the results.

NC STATE UNIVERSITY

Example

- A researcher wishes to test the claim that the average cost of tuition and fees at a four-year public college is greater than $5700. She selects a random sample of 36 four-year public colleges and finds the mean to be $5950. The population standard deviation is $659. Is there evidence to support the claim at a 0.05? Use the *P*-value method.

NC STATE UNIVERSITY

Solution

Step 1 State the hypotheses and identify the claim. H_0: = \$5700 and H_1: $\mu >$ \$5700 (claim).

Step 2 Compute the test value.

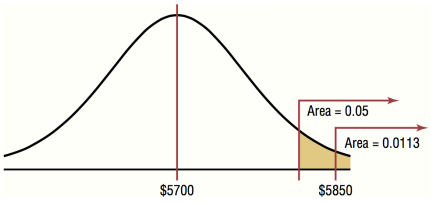
$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{5950 - 5700}{659/\sqrt{36}} = 2.28$$

Step 3 Find the *P*-value. Using Table (here z table) find the corresponding area under the normal distribution for $z = 2.28$. It is 0.9887. Subtract this value for the area from 1.0000 to find the area in the right tail.

$1.0000 - 0.9887 = 0.0113$

Hence the *P*-value is 0.0113.

Step 4 Make the decision. Since the *P*-value is less than 0.05, the decision is to reject the null hypothesis. See Figure below.



Step 5 Summarize the results. There is enough evidence to support the claim that the tuition and fees at four-year public colleges are greater than \$5700.

Note: Had the researcher chosen $\alpha = 0.01$, the null hypothesis would not have been rejected since the *P*-value (0.0113) is greater than 0.01.

NC STATE UNIVERSITY

Important Points About P-Value

- First key difference: The α value is chosen by the researcher before the statistical test is conducted. The P-value is computed after the sample mean has been found.
- However, some researchers do not chose an α value, but report P-value and allow reader to decide whether the null hypothesis should be rejected. In that case, use the following guidelines

Guidelines for *P*-Values

If *P*-value ≤ 0.01, reject the null hypothesis. The difference is highly significant.

If *P*-value > 0.01 but *P*-value ≤ 0.05, reject the null hypothesis. The difference is significant.

If *P*-value > 0.05 but *P*-value ≤ 0.10, consider the consequences of type I error before rejecting the null hypothesis.

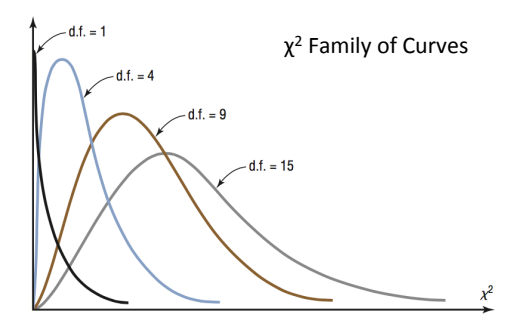
If *P*-value > 0.10, do not reject the null hypothesis. The difference is not significant.

NC STATE UNIVERSITY

Testing for Variance/SD

- Chi-square test for a Variance or Standard Deviation: χ^2
- χ^2 Distribution

The distribution is obtained from the values of $(n-1)s^2/\sigma^2$ when the samples are drawn from a normal distribution with variance σ^2

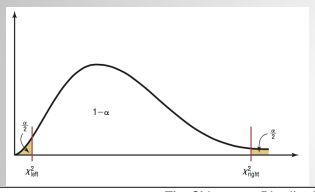


A chi-square variable cannot be negative, and the distributions are skewed to the right. At about 100 degrees of freedom, the chi-square distribution becomes somewhat symmetric. The area under each chi-square distribution is equal to 1.00, or 100%.

NC STATE UNIVERSITY

How to find χ^2 Values?

- Find the values for χ^2_{right} and χ^2_{left} for a 90% c.i. when n = 25?



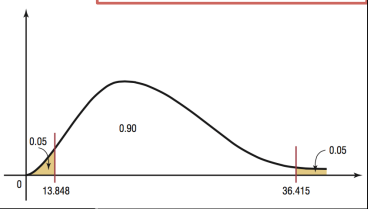
To find χ^2_{right}, subtract 1-0.90=0.10 and divide by 2 to get 0.05.
To find χ^2_{left}, subtract 1-0.05=0.95. Hence, use the 0.95 and 0.05 columns and the row corresponding to 24 d.f.

Two different values are used in the formula because the distribution is not symmetric

The Chi-square Distribution

Degrees of freedom	α 0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1										
2										
⋮										
24							13.848			36.415

χ^2_{left} χ^2_{right}



NC STATE UNIVERSITY

C.I. for a Variance and S.D.

Formula for the Confidence Interval for a Variance

$$\frac{(n-1)s^2}{\chi^2_{\text{right}}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\text{left}}}$$

$$\text{d.f.} = n - 1$$

Formula for the Confidence Interval for a Standard Deviation

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{\text{right}}}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_{\text{left}}}}$$

$$\text{d.f.} = n - 1$$

NC STATE UNIVERSITY

Practice Question

- Find the 95% C.I. for the variance and standard deviation of the nicotine content of a cigarettes manufactured if a sample of 20 cigarettes has a standard deviation of 1.6 milligrams.

NC STATE UNIVERSITY

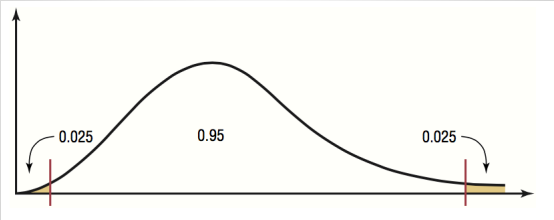
χ^2 test for a Variance and S.D.

- To find area under the chi-square distribution
  - If test is right-tailed, directly get the value at the intersection of α and d.f.
  - If test is left-tailed, then get the value at the intersection of (1-α) and d.f.
  - If the test is two-tailed, note that the area to the right is α/2 and area to the left is (1 – α/2), then find corresponding values from the table

NC STATE UNIVERSITY

Example

- Find the critical chi-square values for n=23 and α=0.05 when a two-tailed test is conducted



What are the α values for right and left?
Then look for corresponding values for d.f. = 22 in the table

NC STATE UNIVERSITY

χ^2 test for a Single Variance

Formula for the Chi-Square Test for a Single Variance

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

with degrees of freedom equal to $n - 1$ and where

n = sample size
s^2 = sample variance
σ^2 = population variance

Assumptions:
- The sample must be randomly selected from the population
- The population must be normally distributed for the variable under study.
- The observations must be independent of one another.

NC STATE UNIVERSITY

Hypothesis testing using χ^2

1. State the hypotheses and identify the claim.
2. Find the critical value(s).
3. Compute the test value.
4. Make the decision.
5. Summarize the results.

NC STATE UNIVERSITY

Testing the Difference Between Two Means

- Suppose one wishes to determine the differences in average GRE scores of engineering graduate students admitted to NCSU and Duke. In other words, does the mean GRE scores of NCSU engineering differs from mean GRE scores of Duke?
- Here the hypotheses are:

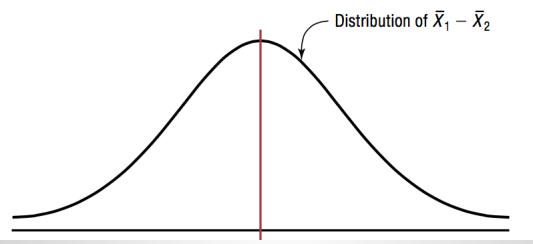
$$H_0: \mu_1 = \mu_2 \quad H_1: \mu_1 \neq \mu_2$$

Alternatively

$$H_0: \mu_1 - \mu_2 = 0 \quad H_1: \mu_1 - \mu_2 \neq 0$$

NC STATE UNIVERSITY

Distribution of $\bar{x}_1$ - $\bar{x}_2$



The variance of the difference is equal to the sum of the individual variances.

$$\sigma^2_{\bar{X}_1 - \bar{X}_2} = \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2}$$

where $\sigma^2_{\bar{X}_1} = \frac{\sigma_1^2}{n_1}$ and $\sigma^2_{\bar{X}_2} = \frac{\sigma_2^2}{n_2}$

So the standard deviation of $\bar{X}_1 - \bar{X}_2$ is

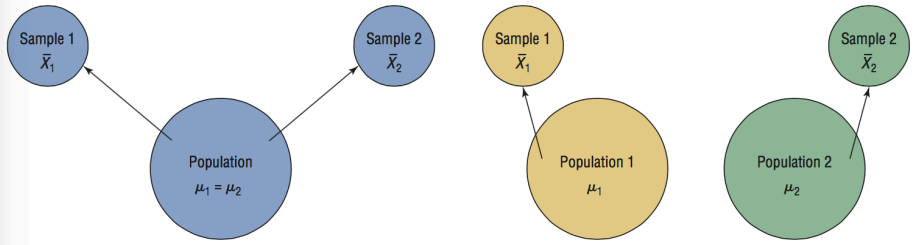
$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Formula for z Test

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

NC STATE UNIVERSITY

Hypothesis-Testing Situations



(a) Difference is not significant

Do not reject H_0: $\mu_1 = \mu_2$ since $\bar{X}_1 - \bar{X}_2$ is not significant.

(b) Difference is significant

Reject H_0: $\mu_1 = \mu_2$ since $\bar{X}_1 - \bar{X}_2$ is significant.

These tests can also be one-tailed, using the following hypotheses:

Right-tailed			Left-tailed		
H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 > \mu_2$	or	H_0: $\mu_1 - \mu_2 = 0$ H_1: $\mu_1 - \mu_2 > 0$	H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 < \mu_2$	or	H_0: $\mu_1 - \mu_2 = 0$ H_1: $\mu_1 - \mu_2 < 0$

NC STATE UNIVERSITY

Basic Format of Hypothesis Testing

1. State the hypotheses and identify the claim.
2. Find the critical value(s).
3. Compute the test value.
4. Make the decision.
5. Summarize the results.

NC STATE UNIVERSITY

Example

- A survey found that the average hotel room rate in Ashville is $88.42 and the average room rate in Raleigh is $80.61. Assume that the data were obtained from two samples of 50 hotels each and that the standard deviations of the populations are $5.62 and $4.83, respectively. At α=0.05, can it be concluded that there is a significant difference in the rates?

NC STATE UNIVERSITY

Solution

Step 1 State the hypotheses and identify the claim.

$H_0: \mu_1 = \mu_2$ and $H_1: \mu_1 \neq \mu_2$ (claim)

Step 2 Find the critical values. Since $\alpha = 0.05$, the critical values are $+1.96$ and -1.96.

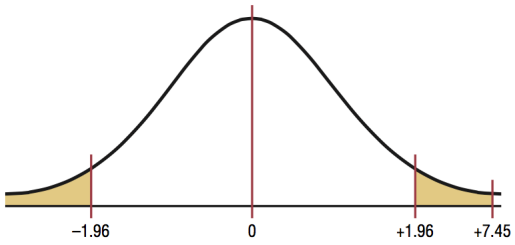
Step 3 Compute the test value.

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(88.42 - 80.61) - 0}{\sqrt{\frac{5.62^2}{50} + \frac{4.83^2}{50}}} = 7.45$$

NC STATE UNIVERSITY

Solution

Step 4 Make the decision. Reject the null hypothesis at $\alpha = 0.05$, since $7.45 > 1.96$.



Step 5 Summarize the results. There is enough evidence to support the claim that the means are not equal. Hence, there is a significant difference in the rates.

NC STATE UNIVERSITY

T Test to Determine the Difference

Variances are assumed to be unequal:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where the degrees of freedom are equal to the smaller of $n_1 - 1$ or $n_2 - 1$.

NC STATE UNIVERSITY

Practice Example

- The average size of a farm in Indiana County, Pennsylvania, is 191 acres. The average size of a farm in Greene County, Pennsylvania, is 199 acres. Assume the data were obtained from two samples with standard deviations of 38 and 12 acres, respectively, and sample sizes of 8 and 10, respectively. Can it be concluded at $\alpha = 0.05$ that the average size of the farms in the two counties is different? Assume the populations are normally distributed.

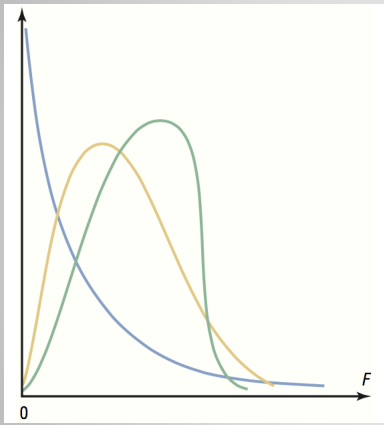
NC STATE UNIVERSITY

Testing the Difference Between Two Variances

- Is there a variation in the temperature for a given month between two different cities?
- Comparing the variance of the cholesterol of men with the variance of cholesterol of women
- For comparing two variances or standard deviations, an *F* test is used.

NC STATE UNIVERSITY

F Distribution



F distribution for different d.f.

If two independent samples are selected from two normally distributed populations in which the variances are equal ($\sigma_1^2 = \sigma_2^2$) and if the variances s^2_1 and s^2_2 are compared as s^2_1 / s^2_2, the sampling distribution of the variances is called the ***F* distribution.**

Characteristics of *F* distribution

- The values of *F* cannot be negative, because variances are always positive or zero.
- The distribution is positively skewed.
- The mean value of *F* is approximately equal to 1.
- The *F* distribution is a family of curves based on the degrees of freedom of the variance of the numerator and the degrees of freedom of the variance of the denominator.

NC STATE UNIVERSITY

Formulae for *F* Test

$$F = \frac{s_1^2}{s_2^2}$$

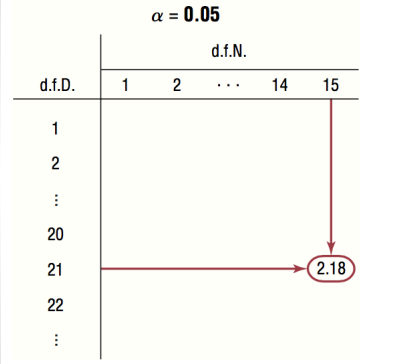
where the larger of the two variances is placed in the numerator regardless of the subscripts.

Note: When you are finding the *F* test value, *the larger of the variances is placed in the numerator of the F formula;* this is not necessarily the variance of the larger of the two sample sizes.

NC STATE UNIVERSITY

Finding C.V.

- Find the critical value for a right-tailed *F* test when α = 0.05, the degrees of freedom for the numerator (abbreviated d.f.N.) are 15, and the degrees of freedom for the denominator (d.f.D.) are 21.



NC STATE UNIVERSITY

Notes and Assumptions

Notes for the Use of the *F* Test

1. The larger variance should always be placed in the numerator of the formula regardless of the subscripts. (See note on page 518.)

$$F = \frac{s_1^2}{s_2^2}$$

2. For a two-tailed test, the α value must be divided by 2 and the critical value placed on the right side of the *F* curve.
3. If the standard deviations instead of the variances are given in the problem, they must be squared for the formula for the *F* test.
4. When the degrees of freedom cannot be found in Table H, the closest value on the smaller side should be used.

Assumptions for Testing the Difference Between Two Variances

1. The populations from which the samples were obtained must be normally distributed. (*Note:* The test should not be used when the distributions depart from normality.)
2. The samples must be independent of each other.

NC STATE UNIVERSITY

Example

- A medical researcher wishes to see whether the variance of the heart rates (in beats per minute) of smokers is different from the variance of heart rates of people who do not smoke. Two samples are selected, and the data are as shown. Using a 0.05, is there enough evidence to support the claim?

Smokers	Nonsmokers
$n_1 = 26$	$n_2 = 18$
$s_1^2 = 36$	$s_2^2 = 10$

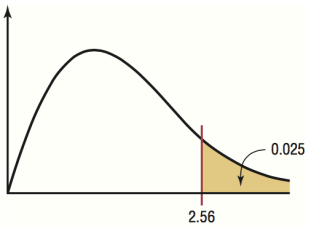
NC STATE UNIVERSITY

Solution

Step 1 State the hypotheses and identify the claim.

$H_0: \sigma_1^2 = \sigma_2^2$ and $H_1: \sigma_1^2 \neq \sigma_2^2$ (claim)

Step 2 Find the critical value. Use the 0.025 table in Table H since $\alpha = 0.05$ and this is a two-tailed test. Here, d.f.N. $= 26 - 1 = 25$, and d.f.D. $= 18 - 1 = 17$. The critical value is 2.56 (d.f.N. $= 24$ was used). See Figure below



Step 3 Compute the test value.

$$F = \frac{s_1^2}{s_2^2} = \frac{36}{10} = 3.6$$

Step 4 Make the decision. Reject the null hypothesis, since $3.6 > 2.56$.

Step 5 Summarize the results. There is enough evidence to support the claim that the variance of the heart rates of smokers and nonsmokers is different.

NC STATE UNIVERSITY

Practice Problems

- Read all definitions, exams includes questions (short answer) on fundamentals
- Read about properties of distributions
  - Questions may include things like:
    - Given frequency plot (figure will be given) identify closest distribution
    - Given various skewed distributions, you should know where widely used statistics lies
    - Given a situation, which probability distribution best applies
- You should know all properties about tests

NC STATE UNIVERSITY

How to compare samples from different populations?

- Height of president L. Johnson is 75in (tallest president) where mean height of president is 71.5 and s.d. is 2.1in. On the other hand height of S. O'Neal's height is 85in (tallest among basket ball players) where mean height of basket ball players is 80in and s.d. is 3.3in. Who is relatively taller?

NC STATE UNIVERSITY

Probability

- Following table summarizes testing 300 subjects for Marijuana usage.

	Did the Subject Actually Use Marijuana? Yes	No
Positive test result (Test indicated that marijuana is *present*.)	119 (true positive)	24 (false positive)
Negative test result (Test indicated that marijuana is *absent*.)	3 (false negative)	154 (true negative)

- Assuming that 1 person is randomly selected from the 300 people that were tested, find the probability of selecting a subject who had a positive test result or used marijuana.
- Assuming that 1 person is randomly selected from the 300 people that were tested, find the probability of selecting a subject who had a negative test result or did not use marijuana.
- Determine whether the following events are disjoint: *A*: Getting a subject with a negative test result; *B*: getting a subject who did not use marijuana.

NC STATE UNIVERSITY

Acknowledgements

- G. James, et. al., Moore, et. al.
- Kleinbaum, et. al.