Statistics Review Session

Christopher Kilner

Basic Math Symbols

Σ (sigma) summation

μ (mu) mean of the population

σ (sigma) standard deviation of population

α (alpha) type I error rate

β (beta) regression coefficients (population)

Θ (theta) a general population parameter

Rounding and Significant Digits

- Non-zero digits are always significant
- Any zeros between two significant digits are significant
- Only a final zero or trailing zeros in the decimal portion are significant
  - It often helps in these cases, or in the case of large numbers, to use scientific notation
  - E.g. 5,000.0101 rounded to 5,000.01*0* simplified to 5.000010×10^3
  - Even in the decimal portion, zeros to the left of the first non-zero digit are not significant
    - E.g. 0.<u>00</u>500
- One frequently used rule of thumb is to round a mean (or a standard deviation) to one additional decimal than the data.

Exponents

- Rules

I. Any number raised to the power of zero equals 1.
   a) $Y^0 = 1$

II. $y^n * y^m = y^{(n+m)}$

III. $(y^n)^m = y^{(m*n)}$

IV. $(y*x)^n = y^n * x^n$

Logarithms

- If $x=b^y$ then $\log_b(x)=y$
- Natural log is ln = log base e, i.e. ($\log_e$)
- **LN(1) is 0**
- **$\text{Log}_b(b) = 1$**

Product Rule:

$$\log_b(MN) = \log_b(M) + \log_b(N)$$

Quotient Rule:

$$\log_b\left(\frac{M}{N}\right) = \log_b(M) - \log_b(N)$$

Power Rule: $\log_b(M^p) = p\log_b(M)$

Factorials

- A factorial, represented by an explanation mark (!), denotes a multiplication of the sequence of descending (natural) numbers.
- Rule
  - N! = N * (N-1)!
  - Example 6! = 6*5*4*3*2*1
  - Shortcut in division:
  - 12!/10! Is the same as 12*11*10!/10! Or 12*11

Algebra

- Review algebra on Calculus Diagnostic Handout
  - https://sites.nicholas.duke.edu/admittedstudents/files/2015/05/Calculus-Review-Handout-1.pdf
  - Both diagnostic exams will cover this topic

Descriptive Statistics

- Sample: *Selected Portion of a Population*
- Population: *Composed of those entities, individuals or objects of interest*
- Qualitative Data
  - Categorical
  - Nominal
- Quantitative Data
  - Discrete
  - Continuous

Descriptive Statistics

Measures of Central Tendency

The Mean

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

The Median

If odd number of observations: middle value (50^{th} percentile)

If even number of observations: halfway between the middle two values

The Mode

The most frequent value.

Descriptive Statistics

Range

largest value minus smallest value

Interquartile Range (IQR)

difference between 75^{th} percentile and 25^{th} percentile values

Variance (s^2)

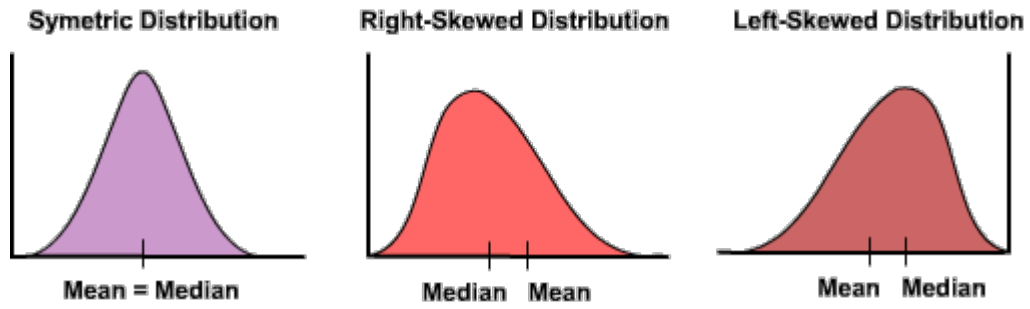
$$\frac{1}{n-1}\sum(x-\bar{x})^2$$

Standard deviation (s)

$$\sqrt{variance}$$

Skew and Outliers

- Skew (description of the distribution of the data)
  - Right-skewed or positive skew occurs when the median is less than the mean
  - Left-skewed or negative skew occurs when the median is greater than the mean



- Outliers:
  - General rule for calculation:
  - 3rd Quartile + 1.5*IQR
  - 1st Quartile - 1.5*IQR

Probability

- A probability ranges from 0 to 1 (or 0% to 100%)
- Therefore, the probabilities of all outcomes must sum to 1, e.g.

$$\sum_{i=0}^{n} p(A_i) = 1$$

- The probability of an event plus the probability of its complement must equal 1.

Probability

- Dependent Events
  - If events are not mutually exclusive, i.e. drawing cards from a deck in succession
  - P(A or/then B) = P(A) + P(B) – P(A and B)
- Independent Events
  - P(A and B) = P(A) * P(B)
- The **law of large numbers** (sometimes named the Law of Averages) states that as the number of trials of a random experiment increases, the empirical probability of an outcome will get closer and closer to its true probability.

Probability Types

- **Marginal probability**: the probability of an event occurring (p(A)), it may be thought of as an unconditional probability. It is not conditioned on another event. Example: the probability that a card drawn is red (p(red) = 0.5). Another example: the probability that a card drawn is a 4 (p(four)=1/13).
- **Joint probability**: p(A and B). The probability of event A **and** event B occurring. It is the probability of the intersection of two or more events. The probability of the intersection of A and B may be written p(A ∩ B). Example: the probability that a card is a four and red =p(four and red) = 2/52=1/26. (There are two red fours in a deck of 52, the 4 of hearts and the 4 of diamonds).
- **Conditional probability**: p(A|B) is the probability of event A occurring, given that event B occurs. Example: given that you drew a red card, what's the probability that it's a four (p(four|red))=2/26=1/13. So out of the 26 red cards (given a red card), there are two fours so 2/26=1/13. *e.g.*

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

Probability Distributions

- Discrete
  - describes a probability distribution of a random variable X, in which X can only take on the values of discrete integers.
  - *Binomial Distribution*: For set of binary independent events, *i.e.* trials, presence/absence, right/wrong, etc…

Binomial distribution

- Fixed number of observation (n)
- Outcome is either success (k) or failure (binary/dichotomous)
- Probability of success (p) is constant across all observations

$$p(k\ successes\ in\ n\ trials) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}\ for\ 0 \le k \le n$$

Probability Distributions

- Continuous: *A probability distribution in which the random variable X can take on any value (is continuous). Because there are infinite values that X could assume, the probability of X taking on any one specific value is zero. Therefore we often speak in ranges of values (p(X>0) = .50).*

Probability Density Function

$$F(x) = P(a \leq x \leq b) = \int_a^b f(x)dx \geq 0$$

- The probability that X falls between two values (a and b) equals the integral (area under the curve) from a to b

The Normal Distribution

- The normal distribution is symmetric and centered on the mean (same as the median and mode). While the x-axis ranges from negative infinity to positive infinity, nearly all of the X values fall within +/- three standard deviations of the mean (99.7% of values), while ~68% are within +/-1 standard deviation and ~95% are within +/- two standard deviations. This is often called the *three sigma rule* or the *68-95-99.7* rule.

Normal Probability Density Function

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \Rightarrow Z = \frac{X - \mu}{\sigma}$$

Student’s t-distribution

- Similar to the normal distribution, the t-distribution is a family of distributions that varies based on the degrees of freedom. A unimodal, continuous distribution, the student’s t distribution has thicker tails than the normal distribution, particularly when the number of degrees of freedom is small. We use the student’s t distribution when comparing means when we do not know the standard deviation of the population and must estimate it from the sample

Sampling

- **Simple random sampling** is one of the core concepts to much of data collection and analysis. *In simple random sampling, each individual or object in a population has an equal probability of being selected into the sample.*
- **Sampling error** is the difference between a sample statistic and the true population parameter
- The standard deviation of the sample means is called the **standard error of the mean**

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

Standard Deviation vs. Standard Error

SE vs. SD

- **What is the difference between standard deviation and standard error?**
  - SD is the typical deviation from the average; it doesn't depend on random sampling.
  - SE is the typical deviation from the expected value in a random sample. SE results from random sampling.

The Central Limit Theorem

- As the sample size increases, the sampling distribution of the sample mean (x) concentrates more and more around μ (the population mean). The shape of the distribution also gets closer and closer to the normal distribution as sample size n increases.
  - Think of this as akin to the law of numbers

Calculating Z-Scores

If a question asks on average $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ This is the Standard Error

If a question asks about a single individual $Z = \frac{X - \mu}{\sigma}$

Calculating Z-Scores

Standard Normal Probabilities

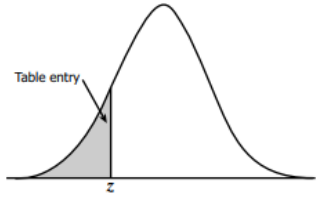


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

- Find probability of a z-score from z-table
- It is important to note that in these tables, the probabilities are the area to the LEFT of the z-score. If you need to find the area to the right of a z-score (Z greater than some value), you need to subtract the value in the table from one.

Inference

- In statistical inference, we take what we know from the sample, apply the underlying theory of sampling (central limit theorem) to make statements about our population of interest.
- Confidence Intervals
  - 95% most common, and assuming CLT:

$$P\left(\mu - 1.96 * \frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + 1.96 * \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Bu
ge $$(\bar{x} - Z^* \frac{\sigma}{\sqrt{n}}, \bar{x} + Z^* \frac{\sigma}{\sqrt{n}})$$

$$P\left(\bar{X} - 1.96 * \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 * \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Confidence Intervals

- Assumptions
  1. We assume the standard deviation of the population (σ) is known.
  2. The sample was randomly selected (independence assumption).
  3. The sample size is large enough to insure that the sampling distribution of the sample means is normally distributed.
  4. There are no outliers (extreme high or low values).

Inference: Comparison of Means

- There are three major types of comparison of means tests:
  1. **One sample test**: We make an inference to a population in comparison to some set value. For example, we might be interest in knowing whether the dissolved oxygen levels in a lake meet a state standard of 5 mg/L.
  2. **Two independent sample test:** In this test, we collect two independent samples to test whether there is a difference in means between two populations (or if one population mean is greater or less than the other) . Comparing GRE scores between men and women is an example of a two independent sample test.
  3. **Paired or Repeated measure test**: This test compares paired data, such as data collected before and after a treatment. Example: a comparison of NOx emissions from randomly selected automobiles before and after an additive is added to the fuel.

Inference: Comparison of Means

- For a comparison of means test, you may use either a **one-sided** or **two-sided** test.
  - A one-sided test (leading to a one-sided p-value) examines whether one mean is greater (or less than) the other mean.
    - If you want to test whether the mean of population A is greater (or less) than the mean of population B, this is a one-sided test.
  - If you want to test whether there is a difference between two means (without any directionality), then you use a two-sided test (and subsequently a tw0-sided p-value (see below).
  - The null and alternative hypotheses should reflect whether or not you are using a one- or two-sided comparison of means test.

Inference: Comparison of Means

- A z-statistic should be calculated when the standard deviation of the population(s) is known.
- If the standard deviation is not known, then the standard error must be estimated using the standard deviation of the sample(s).
  - Due to this estimation, we must use the t-distribution which is thicker in the tails to account for estimating the standard error with the sample standard deviation.

$$\text{test statistic} = \frac{estimate - value\ we\ hypothesize}{standard\ error}$$

$$\text{z-statistic} = \frac{\bar{x} - \mu_o}{\sigma/\sqrt{n}} \qquad \text{t-statistic} = \frac{\bar{x} - \mu_o}{s/\sqrt{n}}$$

Errors

Research

Hypothesis Testing	The Null Hypothesis is True	The Null Hypothesis is False
The Null Hypothesis is True	Accurate	Type II Error
The Null Hypothesis is False	Type I Error	Accurate

Type I and Type II errors

- **Type I error**, also known as a **"false positive"**: the error of rejecting a null hypothesis when it is actually true. In other words, this is the error of accepting an alternative hypothesis (the real hypothesis of interest) when the results can be attributed to chance. Plainly speaking, it occurs when we are observing a difference when in truth there is none (or more specifically - no statistically significant difference). So the probability of making a type I error in a test with rejection region R is $P(R \mid H_0 \text{ is true})$.
- **Type II error**, also known as a "**false negative**": the error of not rejecting a null hypothesis when the alternative hypothesis is the true state of nature. In other words, this is the error of failing to accept an alternative hypothesis when you don't have adequate power. Plainly speaking, it occurs when we are failing to observe a difference when in truth there is one. So the probability of making a type II error in a test with rejection region R is $1 - P(R \mid H_a \text{ is true})$. The power of the test can be $P(R \mid H_a \text{ is true})$.

One Sample t-test

ENV710 Elizabeth A. Albright, PhD

Nicholas School of the Environment

Duke University

Duke | NICHOLAS SCHOOL OF THE ENVIRONMENT

General Steps in Conducting a Comparison of Means Test

1. Decide type of comparison of means test.
   (one sample, two sample, paired samples)

2. Decide whether a one- or two-sided test.

3. Examine the appropriateness of a comparison of means test (based on the assumptions)***

4. Establish null and alternative hypotheses.

5. Decide whether a z-statistic or t-statistic is appropriate.

General Steps in Conducting a Comparison of Means Test

6. Calculate sample mean(s).
7. Calculate standard deviation of sample IF using a t-test.
8. Calculate standard error.
9. Calculate z-statistic or t-statistic.
10. Determine p-value from the test statistic using the appropriate z or t distribution.
11. Interpret the p-value in terms of the hypotheses established prior to the test.

One Sample t-test: Motivating Question

- Do Duke MEM students walk more than 10 miles a week on average?

  One-sided test

- Based on enrollment records, we randomly select 30 full-time, campus-based MEM students and give each a pedometer.

- MEMs wear pedometer and return after a week.

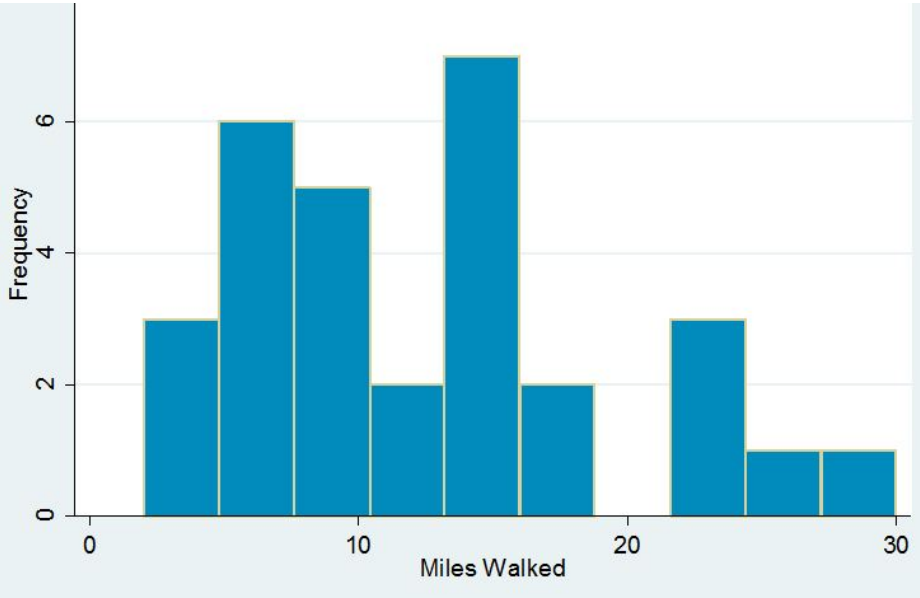
- Establish hypotheses

  Ho: $\mu_{walking} \leq 10$ miles

  Ha: $\mu_{walking} > 10$ miles

Collect the Data

Miles Walked in One Week by MEM Students (n=30)



	Miles
Observations	30
Mean	12.27
Standard Deviation	7.09
Minimum	2
Maximum	30

Assumptions

- Independent observations
  - We randomly selected MEM students to help ensure independence.
- Normally distributed population of miles walked by MEM students
  - Histogram suggests that the population may be roughly normally distributed
  - This assumption becomes more problematic with outliers, heavy skewness and a small sample size.

t-statistic

$$\text{test statistic} = \frac{estimate - value\ we\ hypothesize}{standard\ error}$$

$$\text{t-statistic} = \frac{\bar{x} - \mu_o}{s/\sqrt{n}}$$

t-statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

t-statistic

$$t = \frac{12.27 - \mu_0}{s/\sqrt{n}}$$

t-statistic

$$t = \frac{12.27 - 10}{s/\sqrt{n}}$$

t-statistic

$$t = \frac{12.27 - 10}{7.09/\sqrt{n}}$$

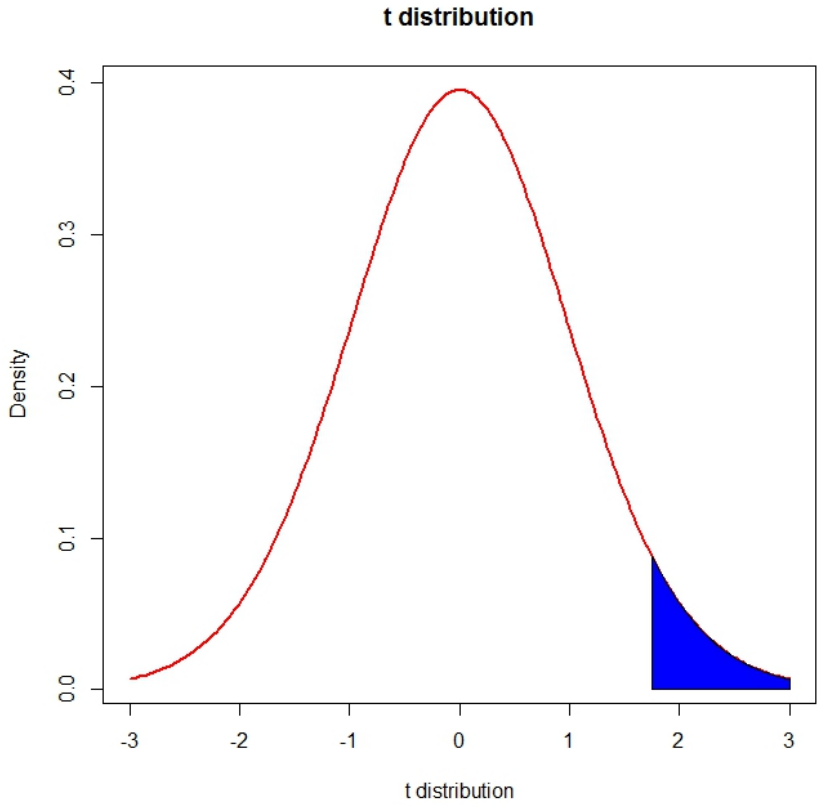
t-statistic

$$t = \frac{12.27 - 10}{7.09/\sqrt{30}}$$

t-statistic

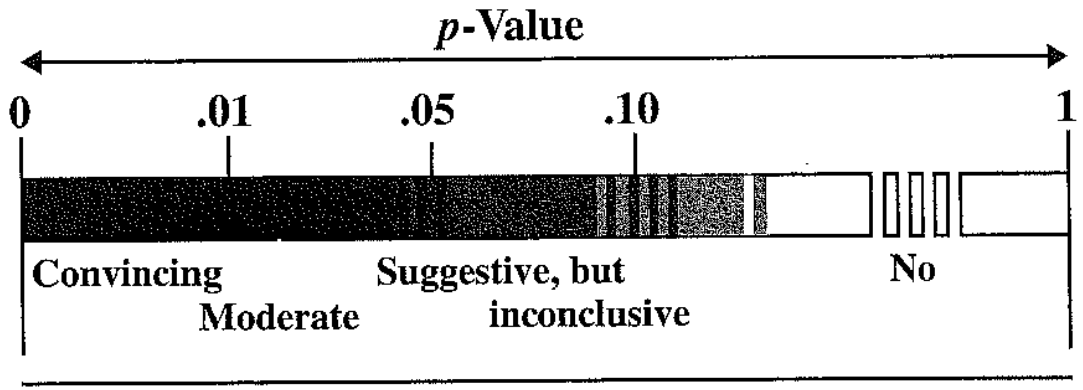
t=1.75, 29 degrees of freedom
p-value = 0.0903

Given that our null hypothesis is true (that Durham residents walk less or equal to than 10 miles/week on average), the probability of getting the results we got, or more extreme is 0.09.



How strong is the evidence?

Display 2.12 Interpreting the size of a *p*-value



- Ramsey and Schafer (2002). The Statistical Sleuth. A Course in Methods of Data Analysis, Second Edition, p. 47.

Conclusion

- Mildly suggestive, but inconclusive, evidence that Durham residents, on average, walk more than 10 miles a week.

Bayes Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})}$$