A NOMAD'S GUIDE TO STATISTICAL METHODS

MOHAMED SAMI ELBIALY

[1] Comments and examples included in these notes are not sufficient for a complete course on elementary differential equations. They do not replace the lecture notes nor the textbook. They are mere elaborations on some of the concepts, ideas, techniques, examples and exercises that are discussed in the text and lectures. You are expected to attend all the lectures and take notes and study them and read the textbook and do all the recommended exercise.

1. Looking at Data Distributions

(1) Individuals and variables.
(2) Categorical variables.
(3) Quantitative variables.
(4) Distribution of a variable: what values the variable takes; frequencies (how often it takes each value); and relative frequencies (percent or fraction.

1.1. Displaying distributions with graphs.

(1) Bar graphs.
(2) Pie chart.
(3) Stemplots.
(4) Back to back stemplots to compare two distributions.
(5) Stemplots with splitting.
(6) Examining a distribution:
   (a) Overall pattern: shape, centre and spread.
   (b) Deviations : outliers.
   (c) Modes: different major peaks of the distribution.
   (d) Unimodal distribution.
   (e) Symmetric and skewed distributions.
(7) Histograms.
(8) Time plots and trends.

1.2. Describing distributions with numbers.

(1) The mean (average).

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i}^{n} x_i$$

(2) The median M.
(3) Mean versus median. Nonresistant versus resistant measures.
(4) Measuring spread:
   (a) The quartiles: Q_1= lowest quarter , Q_3= top quarter .
   (b) The five-number summary: Min, Q_1, M, Q_3, Max.
   (c) Boxplots.
   (d) The p^{th} percentile.
   (e) Interquartile range: $IQR = Q_3 - Q_1$.
(5) The $1.5 \times IQR$ criterion. Call an observation a suspected outlier if it falls $1.5 \times IQR$ above the third quartile Q_3 or below the first quartile Q_1.
(6) Measuring spread: standard deviation.
   (a) The variance s^2:

$$s^2 = \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

   (b) the standard deviation s

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

   (c) A few questions:
      (i) Why don't we just add up the deviations

$$x_1 - \bar{x}, x_2 - \bar{x}, \cdots, x_n - \bar{x}?$$

      (ii) Why do we square the deviations and not just add up

Date: November 29, 2005.
[1]Mathematics Department
University of Toledo
Toledo, Ohio 43606
U.S.A.

the distances

$$|x_1 - \bar{x}|, |x_2 - \bar{x}|, \cdots, |x_n - \bar{x}|?$$

(iii) Why do we emphasize the standard deviation s rather than the variance s^2?
(iv) Why do we average by dividing $n-1$ and not by n?
(v) Is the standard deviation a resistant or a nonresistant measure? Why?

(d) Another question: When is the *five number summary* better for describing a distribution and when are the *mean and standard deviation* better?

1.3. **The normal distributions.**

(1) **Density curve.** Fill in the space:
(a) The ..... of a density curve is the balance point, at which the curve would balance if it were made of solid material. Think about observations as weights hanging on a weightless rod.
(b) The ..... of a density curve is the point that divides the area under the curve in half. This means half the observation lie on the left and the other half lie on the right.
(c) Normal curves: bell shaped, symmetric and unimodal. Median and mean ..... .

(2) **The normal distribution with mean μ and standard deviation σ.**
(a) The density function of $N(\mu, \sigma)$ is

$$\boxed{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}}$$

(b) When does a normal curve changes concavity?
(c) The $68 - 95 - 99.7$ rule.

(3) **The standard normal distribution $N(0, 1)$.**
(a) If a variable X has any normal distribution $N(\mu, \sigma)$, then the *standardized variable*

$$\boxed{Z = \frac{X - \mu}{\sigma}}$$

has the standard normal distribution.
(b) The standardized value of an observation x is given by

$$\boxed{x = \frac{x - \mu}{\sigma}}$$

and is called a z-score.

(4) **Normal quantile plots** AKA (normal probability plots) for a distribution $f(x)$.
(a) In a normal quantile plot, $x = H(z)$, where given z, x is the unique value such that

$$\boxed{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}}\, dt = \int_{-\infty}^{x} f(t)\, dt}$$

(b) In practice, we are given the distribution of a discrete variable X in a form of a table. How do you graph a normal quantile plot.
(c) How do you decide whether the distribution of a given data is approximately a normal distribution?
(d) How do you decide whether the distribution of a given data is approximately a standard normal distribution?
(e) How do you decide whether the distribution of a given data is right-skewed or left-skewed? Be careful which variable is on the horizontal axis and which is on the vertical axis.
(f) How do you identify outliers in a normal quantile plot?
(g) Granularity.

2. Looking at Data Relationships

(1) Association between variables.
(2) Response variable, explanatory variable.
(3) Independent variables and dependent variables.

2.1. Scatterplots.

(1) Examining a scatterplot: overall pattern (form, direction, strength), deviation from overall pattern, outliers.
(2) Which do you plot on the horizontal axis, the response variable or the explanatory one? Which do you plot on the vertical axis?
(3) Clusters.
(4) Positive association, negative association.
(5) Adding categorical variables to scatterplots.

2.2. Correlation.

$$\boxed{r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)}$$

Notice that in computing the correlation r we standardize the variable X and Y.

Properties of correlation:

(1) How does *correlation* distinguish between the explanatory variable and the response variable?
(2) Which of the following does the correlation require?
  (a) Both variables to be qualitative.
  (b) One variable to be qualitative and one variable to be quantitative.
  (c) Both variables to be quantitative.
(3) How does correlation r change with the change in units of x and y? Why?
(4) Positive r indicates .....
(5) Negative r indicates .....
(6) r close to 0 indicates .....
(7) The correlation r satisfies : $\cdots\cdots \leq r \leq \cdots\cdots$
(8) $r = \pm 1$ means .....
(9) What kind of relationship does the correlation measure?
(10) Is the correlation a resistant or a nonresistant measure?

2.3. Least square regression.

(1) What is a regression line?
(2) What is a regression line used for?
(3) Equation of the least-squares regression line of a response variable y on an explanatory variable x is

$$\boxed{\hat{y} = a + bx}$$

with slope

$$b = \cdots \frac{\cdots}{\cdots}$$

and intercept

$$a = \cdots\cdots$$

(4) Interpreting the regression line:
  (a) The expression $b = \cdots$ for the slope says that along the regression line, a change of one standard deviation in x corresponds to a change of $\cdots$ standard deviations in $\cdots$.
  (b) The least-square regression line always passes through the point $(\cdots, \cdots)$.
  Explain.
  (c) If both X and Y are standardized variables, then the regression line has a slope $\cdots$.
(5) Correlation and regression:
  (a) Recall that the correlation treats X and Y equally.
  (b) Least-square regression depends on the distance of the data points from the line only in the y direction. This is why the explanatory variable X and response variable Y play different roles in regression. this means that the regression line of y on x is different from the regression line of x on y.
  (c) What is the equation of the regression line of x on y
(6) r^2 in regression: r^2 is the fraction of the variation in the values of y that is explained by the least regression of y in x. Moreover

$$\boxed{r^2 = \frac{\text{variance of predicted values } \hat{y}}{\text{variance of observed value } y}}$$

2.4. **Cautions about regression and correlation.**

(1) Residuals.
(2) Residual plots:
    (a) Unstructured horizontal band .....
    (b) A curved pattern .....
    (c) A fan-shaped pattern .....
(3) What is the mean of the least-square residual? Explain.
(4) Lurking variables.
(5) What is the difference between outliers and influential observations?

2.5. **The question of causation.**

(1) Causation.
    - x = amount of saccharin in a rat's diet.
    - y = count in tumor in rat's bladder.
(2) Common response.
    - x = a student's SAT score.
    - y = the student's 1^{st} year gpa.
(3) Confounding.
    - x = number of years of education.
    - y = income.
(4) What about
    - x = mother's body mass index (weight relative to height)
    - y = daughter's body mass index.

2.6. **Transforming relationships.**

3. Producing Data

4. Probability-the Study of Randomness

4.1. **Randomness.**

(1) Random phenomenon.
(2) Probability.

4.2. **Probability models.**

(1) Sample space.
(2) Event.
(3) Probability rules. Explain the following rules in words:
  (a) $0 \leq P(A) \leq 1$.
  (b) $P(S) = 1$.
  (c) $P(A^c) = 1 - P(A)$.
  (d) $P(A \text{ or } B) = P(A) + P(B)$ provided that A and B are disjoint.
(4) Venn diagrams.
(5) Probabilities in a finite sample.
(6) Equally likely outcomes.
(7) Independence.
(8) Multiplication rule for independent events.

4.3. **Random variables.**

(1) Discrete random variable X.
  (a) Probability distribution of a discrete random variable X.

Values of X	x_1	x_2	x_3	$\cdots$	x_n
$P(X = x_j)$	p_1	p_2	p_3	$\cdots$	p_n

    (i) Each $0 \leq p_j \leq 1$.
    (ii) $p_1 + p_2 + \cdots + p_n = 1$.
    (iii) How do you find the probability of the event $\{x_2, x_5, x_{12}, x_{23}\}$? That is, how do you find $P(X = x_2, x_5, x_{12} \text{ or } x_{23})$?
  (b) Probability histogram for discrete random variable.
(2) Continuous random variable X, $a < x < b$.
  (a) The probability distribution is given by a density curve.
  (b) Area under the curve and above the interval $[a, b] = \cdots$.
  (c) Area under the curve and above the interval $[a_1, a_2] = \cdots$
  (d) Probability of an individual out come x_o is $P(X = x_o) = \cdots$.

4.4. **Means and variances of random variables.**

(1) The mean of a random variable .
  (a) The mean of a random variable X is also called *the expected value of* X.
  (b) The mean of a discrete random variable X is a weighted average of all possible outcomes of the the random variable X. In this average, each outcome x_i is weighted by its probability

$$p_i = \frac{\#(\text{ occurrences of } x_i)}{\text{total number of all possible outcomes}} = \frac{\#(x_i)}{\sum_{j=1}^{n} \#(x_j)}$$

Therefore,

$$\boxed{\mu_X = \sum_{i=1}^{n} x_i p_i}$$

  (c) What is the difference between μ_X and $\bar{x}$?
(2) Simple random sample , SRS.
(3) Statistical estimation and the law of large numbers.
  (a) μ is called a *parameter* and $\bar{x}$ is called a *statistic*. What is the difference?
  (b) The law of large numbers.
  (c) "The law of small numbers".
(4) Rules for means. Let X and Y be two random variables.

$$\boxed{\mu_{a+bX} = a + b\mu_X}$$

$$\boxed{\mu_{X+Y} = \mu_X + \mu_Y}$$

(5) The variance of a random variable.

$$\boxed{\sigma_X^2 = \sum_{i=1}^{n} (x_i - \mu_X)^2 p_i}$$

The standard deviation is

$$\boxed{\sigma_X = \sqrt{\sigma_X^2}}$$

(6) Rules for variances.

$$\boxed{\sigma_{a+bX}^2 = b^2 \sigma_X^2}$$

Two random variables X and Y are independent if the correlation between them is zero.

If X and Y are independent random variables, then
$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$$
$$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$$

If X and Y have correlation ρ, then
$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y + 2\rho\sigma_X\sigma_Y$$
$$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y - 2\rho\sigma_X\sigma_Y$$

4.5. General probability rules.

(1) Rules of probabilities.
(2) Addition rules for
   (a) unions of disjoint events,
   (b) unions of non-disjoint event.
(3) Conditional probability. Intersection of two event:

$$P(A\&B) = P(A)P(B|A)$$

$$P(B|A) = \frac{P(A\&B)}{P(A)}, \qquad \text{when } P(A) > 0$$

If A and B are independent $P(A\&B) = P(A)P(B)$. Thus,

$$P(B|A) = P(B), \quad \text{for independent A and B}$$

(4) Multiplication rule. Intersection of three events:

$$P(A\&B\&C) = P(A)P(B|A)P(C|A\&B)$$

(5) Intersection.

5. Sampling Distributions

Introduction.

(1) Statistical inference draws conclusions about a ..... or a ..... on the basis of ..... .
(2) Sampling distribution verses population distribution.

5.1. Sampling distributions for counts and proportions.

(1) Count verses sample proportion : X verses $\hat{p} = X/n$.
Recall that from the "rules for mens and standard deviations" above we have

$$\mu_{\hat{p}} = \frac{1}{n}\mu_X, \qquad \sigma^2_{\hat{p}} = \frac{1}{n^2}\sigma^2_X$$

(2) Binomial distribution with parameters n and p called $B(b, p)$:
   (a) There are n independent observations.
   (b) Each observation is either success or failure.
   (c) p = probability of success in each observation.
   (d) $0 \le X \le n$ is the number of success. X is a nonnegative integer.
(3) The binomial probability.

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

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

(4) Examples. Which is a binomial setting? Explain and give n and p if it is:
   (a) Tossing a balanced coin 10 times.
   (b) Tossing and unbalanced coin 10 times.
   (c) Dealing 10 cards from a shuffled deck and count the number of black cards.
   (d) Reliability.
(5) When do we use the binomial sampling distribution to make to make inferences about the proportion p of success?

(6) Binomial probability table C.
(7) Binomial mean and standard deviation.

$$\mu_X = np$$
$$\sigma_X = \sqrt{np(1-p)}$$

(8) $\hat{p}$ = sample proportion of success in an SRS of size n.

$$\hat{p} = \frac{\text{count of success in sample}}{\text{size of sample}} = \frac{X}{n}$$

(9) Important remark. The proportion $\hat{p}$ does not have a binomial distribution because it is not a count.
(10) Mean and standard deviation of the proportion $\hat{p}$.

$$\mu_{\hat{p}} = \frac{1}{n}\mu_X = p$$
$$\sigma^2_{\hat{p}} = \frac{1}{n^2}\sigma^2_X$$

Thus,

$$\mu_{\hat{p}} = p$$
$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

(11) Normal approximation for counts and proportions when n is large.
  (a) By "n large" it is meant that $np \geq 10$ and $n(1-p) \geq 10$ in the same time. That is, both the mean of the count of success and and the mean of the count of failure are ≥ 10.
  Question: does this mean that it is enough to take $n \geq 20$?
  (b) In this case, X is approximated by $N(np, \sqrt{np(1-p)})$.
  (c) And $\hat{p}$ is approximated by

$$N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

What does this all mean? It means that
  (a) if we choose n large such that

$$np \geq 10 \text{ and } n(1-p) \geq 10$$

(for example if $p = 0.3$ we want $n \geq 10/p = 100/3 = 33.33333$ and $n \geq 10/0.7 = 14.29$. In this case $n \geq 34$ suffices. Let's take $n = 40$.)
  (b) then we draw an SRS of size n a large number of times,
  (c) and each time we counted the number of successes X and computed $\hat{p} = X/n$
  (d) Then if we plot the histogram of all the $\hat{p}$'s, we get a normal distribution.

$$N\left(p, \sqrt{\frac{p(1-p)}{n}}\right) = N\left(0.3, \sqrt{\frac{0.3(0.7)}{40}}\right)$$

  (e) And if we plot the histogram of all the X's, we get a normal distribution

$$N(np, \sqrt{np(1-p)}) = N(40(0.3), \sqrt{40(0.3)(0.7)})$$

5.2. **The sampling distribution of a sample mean $\bar{x}$.**

- Counts and proportions are discrete random variables that describe categorical data: X is the number of "yes's" and $\hat{p}$ is their proportion.
- Sample means, percentiles, and standard deviations describe quantitative data. They are continuous random variables.

What do we mean by the sampling distribution of a sample mean $\bar{x}$?
Suppose we are interested in the average height of our university students.

- We can measure the height of the all the 20,000 and compute the exact mean μ and standard deviation σ.
- Or we can pick an SRS of 10 students at random and measure their height and compute the sample mean $\bar{x}$.
- What happens if we repeat this experiment a large number of times? That is, what happens if pick a large number, say a hundred of SRS's of size

$n = 10$ and compute their sample means $\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_{100}$.

- What happens is that $\bar{x}$ becomes a random variable of its own right, with its own mean $\mu_{\bar{x}}$ and standard deviation $\sigma_{\bar{x}}$.
- The sampling distribution of a sampling mean is the distribution of the $\bar{x}$'s.
- If we pick all possible SRS's of size 10, then the distribution of $\bar{x}$ is normal. However, this would be very large number of samples $\begin{pmatrix} 20,000 \\ 10 \end{pmatrix}$.
- But if we pick a smaller number of SRS, 100 for example, then the distribution of $\bar{x}$ is close to normal.
- In order to understand what's going on, we need to find $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$ in terms of μ and σ.

The mean and standard deviation of $\bar{x}$. Now each individual in the SRS is chosen at random. So, we have n random variables $X_1, X_2, \cdots$, and X_n, one for each individual. Each of these X's has mean and standard deviation of the entire population, that is μ and σ respectively.

$$\bar{x} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

From our knowledge of probabilty, we have

$$\mu_{\bar{x}} = \mu$$
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Important observation: The spread of $\bar{x}$ is less then The spread of X of the population, because $\sigma = \sigma_X$ and

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} < \sigma$$

This means that averaging over more measurements reduces variability and makes it more likely that $\bar{x}$ is close to μ.

If a population has the $N(\mu, \sigma)$ distribution, then the sample mean $\bar{x}$ of n independent observation has the $N(\mu, \sigma/\sqrt{n})$ distribution.

What if the population does not have a normal distribution? In this case we have the central limit theorem:

Central limit theorem.
Draw an SRS of size n from any population with mean μ and standard deviation σ. When n is large, the sampling distribution of the sample mean $\bar{x}$ is approximately the normal distribution $N(\mu, \sigma/\sqrt{n})$.

6. Introduction to Inference

In this chapter we study *confidence intervals* and *tests of significance.*
Important assumption. We assume that the data comes from a random sample or a randomized experiment.

6.1. Estimating with confidence.

(1) We want to estimate μ of the population from data $x_1, \cdots, x_n$.
(2) We assume that we know σ of the population.
(3) We require confidence level C, for example $C = 0.95, 0.97, \cdots$. That is we want to be 95% confident that $\bar{x}$ that we compute is close enough to the actual μ so that its z-score satisfies

$$\frac{|\bar{x} - \mu|}{\sigma/\sqrt{n}} \leq z_c$$

Where

$$P(Z \geq z_c) = \frac{\alpha}{2} = \frac{1-C}{2} = \frac{0.5}{2}, \frac{0.3}{2}, \cdots$$

(4) In this case the confidence interval is

$$\boxed{\text{(a)} \qquad \bar{x} - z_c\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_c\frac{\sigma}{\sqrt{n}}}$$

(5) Explain (a) with a graph.

(6) We also write the confidence interval as

$$\boxed{\text{(b)} \qquad \bar{x} \pm z_c\frac{\sigma}{\sqrt{n}}}$$

(7) Margin of error.
(8) Suppose you increase your level of confidence C from 95% t0 97%, what happens to the confidence interval? Is it larger or smaller? Is this higher level of confidence better? Why?

Example Suppose that repeated measurements of a certain quantity A vary normally with $\sigma = 0.3$.

(1) One measurement is made and the result is the result is $x = 4.6$. Give a 90% confidence interval for the mean.
Answer: Find z_c such that $P(Z \geq z_c) = (1 - 0.9)/2$. $z_c = 1.645$.
Notice that in this case $n = 1$
The 90% confidence interval for the mean is

$$4.6 \pm (1.645)(0.3)/\sqrt{1} = 4.6 \pm 0.4935$$

$$4.1065 \leq \mu \leq 5.0935$$

(2) Four measurements are made and the average is $\bar{x} = 4.6$. Give a 90% confidence interval for the mean.
Answer: We use the same $z_c = 1.645$.
Notice that in this case $n = 4$
The 90% confidence interval for the mean is

$$4.6 \pm (1.645)(0.3)/\sqrt{4} = 4.6 \pm 0.2467$$

I.e.

$$4.3433 \leq \mu \leq 4.8467$$

(3) What does this mean?

Definition We call $\sigma/\sqrt{n}$ the *standard error*,

$$SE = \frac{\sigma}{\sqrt{n}}$$

We can write the confidence interval as

$$\boxed{\text{(c)} \qquad \bar{x} - z_c SE \leq \mu \leq \bar{x} + z_c SE}$$

6.2. Tests of significance.

What is the difference between *confidence interval* and test of significance?
Confidence interval is used when we are trying to estimate a population parameter such as the population mean μ.
Tests of significance are used to measure the extent to which the collected data constitute an evidence against a certain hypothesis

Example 1.

(1) Suppose the mean and standard deviation of a population was measured a while ago and found to be $\mu = 199$ and $\sigma = 6$. We know that for this population the standard deviation does not change.

However, we predict that the mean should have increased since the last time it was measured.

(2) We make four measurements of $x_1 = 200, x_2 = 204, x_3 = 206, x_4 = 2010$. The sample mean is $\bar{x} = 205$.

(3) The question now is: to what extend does the collected data constitute an evidence against the hypothesis that (μ hasn't increased)?

(4) Notice that the claim we are testing (μ hasn't increased) is the opposite of what we predict (μ has increased.)

(5) Formulating the question: We call the claim that we want to refute (μ hasn't increased) *the null hypothesis* H_o.
We call the alternative claim that we predict *the alternative hypothesis* H_a.

$$H_o : \mu = 199, \qquad H_a : \mu > 199$$

The question can be phrased as follows:

If H_o is true,
what is the probability that $\bar{x} \geq 205$?

Question: What does it mean if this probability is small? And what does it mean if this probability is?
Answer: If this probability is is small, then null hypothesis H_o is unlikely and our prediction (the alternative hypothesis H_a is more likely.

(6) Finding the sought probability:
The standardized sample mean is

$$z = \frac{205 - 199}{6/\sqrt{4}} = 2$$

z is called the test statistic.

$$P(\bar{x} \geq 205) = P(Z \geq 2) = 0.0228$$

This tells us that,

if $\mu = 199$
the probability that we obtain
a sample mean $\bar{x} = 205$
(i.e. obtain these four measurements)
is 0.023

In other words, if we keep repeating the experiment (of taking samples of four and compute the average time and again) we will get $\bar{x} \geq 205$ no more than 2.3% of the time.

The P-value of test of significance.

(1) The probability that computed (i.e. *the probability that assuming H_o is true, a test statistic is $\geq$ the measured one*) is called the P-value of the test.

(2) The smaller the P-value, the stronger the evidence against H_o that the data provides.

One sided and two sided alternative hypothesis H_a for μ.

one-sides:
$H_o : \mu = \mu_o, \quad H_a : \mu > \mu_o$
with P-value:
$$P(Z \geq \frac{\bar{x} - \mu_o}{\sigma/\sqrt{n}})$$

two-sides:
$H_o : \mu = \mu_o, \quad H_a : \mu \neq \mu_o$
with P-value:
$$P(|Z| \geq \frac{|\bar{x} - \mu_o|}{\sigma/\sqrt{n}}) = 2P(Z \geq \frac{|\bar{x} - \mu_o|}{\sigma/\sqrt{n}})$$

Example 2. If we have a two-sided alternative hypothesis for the previous example

$$H_o : \mu = 199, \qquad H_a : \mu \neq 199$$

then the P value is

$$\begin{aligned} P(|Z| \geq \frac{|\bar{x} - \mu_o|}{\sigma/\sqrt{n}}) &= P(|Z| \geq 2) \\ &= 2P(Z \geq 2) \\ &= 2(0.0228) = 0.0456 \end{aligned}$$

Significance level: How much evidence against H_o is enough to reject H_o?

(1) **Significance level** α means that we reject H_o if the P-value is less than α.

(2) In this case we say that the data is statistically significant at level α.

Example 3.

(1) If $\alpha = 0.05$,
   (a) We reject H_o in example 1 at the 0.04 significance level because $p = 0.0228 < 0.05$.
   (b) We reject H_o in example 2 at the 0.04 significance level because $p = 0.0456 < 0.05$.
(2) If $\alpha = 0.04$,
   (a) We reject H_o in example 1 at the 0.04 significance level because $p = 0.0228 < 0.04$.
   (b) We cannot reject H_o in example 2 at the 0.04 significance level because $p = 0.0456 > 0.04$.
(3) If $\alpha = 0.02$,
   (a) We cannot reject H_o in example 1 at the 0.04 significance level because $p = 0.0228 > 0.02$.
   (b) We cannot reject H_o in example 2 at the 0.04 significance level because $p = 0.0456 > 0.02$.

Two sided significance test and confidence interval.

(1) Recall that the confidence interval is an interval that would include the true value of μ with probability C. For example $C = 0.90, 0.95, 0.97, \cdots$.
(2) Suppose we have

$$H_o : \mu = \mu_o, \qquad H_a : \mu \neq \mu_o$$

(3) A level α two-side significance test rejects H_o iff the value μ_o falls outside the the $c = 1 - \alpha$ confidence interval for μ.
(4) To see this recall that for the confidence interval we find z_c such that

$$P(Z \geq z_c) = \frac{1 - C}{2}$$

and the confidence interval is

$$\bar{x} \pm z_c \frac{\sigma}{\sqrt{n}}$$

We reject H_o at a level α two-side significance test if we compute

$$z_* = \frac{\bar{x} - \mu_o}{\sigma/\sqrt{n}}$$

and find out that

$$2P(Z \geq z_*) < \alpha$$

In either case

$$z_c = z_* = \frac{\bar{x} - \mu_o}{\sigma/\sqrt{n}}$$

7. Inference For Distributions

7.1. **Inference for a mean of a population.** So far we have been assuming that we know the standard deviation of the population σ. If we don't, we estimate it from the data.

(1) Compute both the mean $\bar{x}$ and standard deviation s of the data.
(2) Define the standard error

$$\boxed{SE_{\bar{x}} = \frac{s}{\sqrt{n}}}$$

(3) Instead of using the one-sample z statistic

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

we use the t distribution with $n-1$ degrees of freedom.
(4) Suppose that an SRS of size n is drawn from an $N(\mu, \sigma)$ population

The one-sample t statistic

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

has the t distribution with $n-1$ degrees of freedom.

(5) We proceed as before to compute confidence intervals and perform tests of significance, except that now we use table D for the t distribution.
(6) The density curve of the t distribution is also bell-shaped and symmetric around 0, but it is flatter then the z distribution. That is, it has more area (probability) near the tails.
(7) **The one-sample confidence interval.**

$$\bar{x} \pm t_c \frac{s}{\sqrt{n}}$$

$$P(T_{n-1} \geq t_c) = \frac{1-C}{2}$$

(8) **The one-sample t test.**

one-sides:
$H_o : \mu = \mu_o, \quad H_a : \mu > \mu_o$
with P-value:
$P(T_{n-1} \geq t)$

one-sides:
$H_o : \mu = \mu_o, \quad H_a : \mu < \mu_o$
with P-value:
$P(T_{n-1} \leq t)$

two-sides:
$H_o : \mu = \mu_o, \quad H_a : \mu \neq \mu_o$
with P-value:
$P(|T_{n-1}| \geq t)$
$= 2P(T_{n-1} \geq t)$

7.2. **Comparing two means.** To compare the responses in two groups. For example when testing a new medicine. One group is given the medicine and the other is given a placebo.

(1) Each group is a sample from a distinct population.
(2) The responses in each group are independent of those in the other group.
(3) We want to estimate $\mu = \mu_1 - \mu_2$.
(4) We define a new random variable $Y = X_1 - X_2$.
(5) From the addition laws of probability, we have

$$\mu_Y = \mu = \mu_1 - \mu_2$$
$$\bar{y} = \bar{x}_1 - \bar{x}_2$$
$$\sigma_{\bar{y}}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$
$$s_{\bar{y}}^2 = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$$

(6) **The two sample z-statistic.**

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

(7) **The two sample t-statistic.**

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

(8) **Degrees of freedom for the t statistic:** either approximated by software, or we use tables D with

$$\boxed{k = df = \min\{n_1 - 1, n_2 - 1\}}$$

with degrees of freedom k.

(9) The t statistic has *approximatly* the $t(k)$ distribution.

(10) **The two sample significance test.** We want to test the null hypothesis

$$H_o : \mu_1 = \mu_2$$

This is equivalent to the null hypothesis

$$H_o : \mu_1 - \mu_2 = 0$$

(11) If we know σ_1 and σ_2, we use z significance test

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

(12) If we do not know σ_1 and σ_2, we use t significance test

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

(13) **The two-sample t test.**

$$\boxed{\begin{array}{c}\text{one-sides:}\\ H_o : \mu_1 = \mu_2, \quad H_a : \mu_1 > \mu_2\\ \text{with P-value:}\\ P(T_k \geq t)\end{array}}$$

$$\boxed{\begin{array}{c}\text{one-sides:}\\ H_o : \mu_1 = \mu_2, \quad H_a : \mu_1 < \mu_2\\ \text{with P-value:}\\ P(T_k \leq t)\end{array}}$$

$$\boxed{\begin{array}{c}\text{two-sides:}\\ H_o : \mu_1 = \mu_2, \quad H_a : \mu_1 \neq \mu_2\\ \text{with P-value:}\\ P(|T_k| \geq t)\\ = 2P(T_k \geq t)\end{array}}$$

(14) **The two sample t confidence interval.**

$$\boxed{(\bar{x}_1 - \bar{x}_2) \pm t_c \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where the area under the $t(k)$ density curve, above the interval $[-t_c, t_c]$ is C.

(15) **Pooled two-sample t procedures.**

(a) We use this procedure when $\sigma_1 = \sigma_2$.

(b) In this case the t statistic has exactly the t distribution.

(c) The estimator for $\sigma^2 = \sigma_1^2 = \sigma_2^2$ is

$$\boxed{s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

(d) A closer look at s_p^2 shows that it is the weighted average of s_1^2 and s_2^2 with weights equal to the degrees of freedom. This gives more weight to the larger sample since it provides us with more information. Recall that when we compute standard deviation we average by $n - 1$ and not by n. Do you remember why we do that?

(e) Now what is $s_{\bar{y}}$?
Recall that $\sigma_1^2 = \sigma_2^2$. Thus

$$\boxed{s_{\bar{y}}^2 = \frac{s_p^2}{n_1} + \frac{s_p^2}{n_2} = s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}$$

Thus

$$\boxed{s_{\bar{y}} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

(f) The t statistic is

$$\boxed{t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}$$

with degrees of freedom

$$\boxed{k = n_1 + n_2 - 2}$$

Then we use t as above for significance tests and confidence intervals.

MATH 3610
Statistical Methods I
Fall 2004
ElBialy
Exercises

(1) The following problems are intended to show the types of methods and techniques and ideas that we cover in each section.
(2) Problems in each group are similar. Most of the time, questions in the same group ask you the same question about a number of objects, often functions.
(3) You should do as many problems as you need to master the ideas and techniques and methods we cover in each section. Doing that, will be very helpful to you on exams, quizzes team work etc. ... However, you need also to study the material we cover in the text and your lecture notes.

§1.1: 14, (22,43) (24,44), 28, 31(2.14, 2.30, 2.56), 25, 26, 30, 31, 33, 36, 39.
§1.2: 41, (43, 22), (44, 24), (46,33) , 47, 48,49, 55-57, 62-64.
§1.3: 87-91, 92-95, 108, 109, 110.
Ch1 Ex: 119, 123, 128, 129, 130, 131, 133, 134 , 135, 136.
: **********
§2.1: 3, 6, (9,47), (11, 27, 42), (53, 13).
§2.2: 23, 24, (11, 27).
§2.3: (42, 11, 27), (53, 13).
§2.4: 63, 69, 71, 73, 78.
§2.5: 79, 80, 81.
: **********
§4.1: 6, 7, 8.
§4.2: 13, 14, 15, 17, 21, 26, 29.
§4.3: (43, 45, 61, 63), 47, 49, 50, 53, (56, 62).
§4.4: 59, (43, 45, 61, 63), (56, 62), 65, 73, 81.
§4.5: 89, 90, 81, (92, 95), (96, 97), (98,99), 103, 104, 105. **********
§5.1: 5, 11, 13, 15, 19, 21, 23.
§5.2: 29,31, 36, 37, 39, 41, 45, 46, 47, 48, 49.
: **********
§6.1: 1, 2, (5,7), 15, (16-18), (20,21), (5, 7, 25,26), 29.
§6.2: 35, 37, 39, 41, 43, 45, 47, 52, 53, 57, 62.
§7.1: 1, 11 13, (14, 15), 16, 18, 33 .
§7.2: 53, 57, 61, 71, ...
: **********
: **********
: **********
: **********

E-mail address: melbialy@math.utoledo.edu