Lecture 8

Hypothesis Testing

Statistics for
Civil & Environmental Engineers

Relax a little and think about issue today!

- **Making Decisions**
  - **Do materials meet specifications?**
  - **Have pollutant levels increased?**
  - **Has streamflow been affected by urbanization?**
  - **Does new blend result in greater strength concrete?**
  - **Does SO_2 affect human health?**
  - **Is acid rain causing environmental damage?**
  - **Have new management procedures improved production?**

How do we organize information?

How much evidence do we demand?

Procedure for Testing

1. Declare a ***null hypothesis*** which is the hypothesis to be tested
2. An ***alternative hypothesis*** which we really wish to test
3. Determine a ***test statistic***
4. Determine a ***level of significance α*** based on the known distribution of the test statistics
5. Define a ***rejection*** (or ***critical***) ***region***
6. Use the observed data to verify whether the computed value of the test statistic is within or outside the rejection region

Probabilities of Type I and Type II Errors(1)

- **Definition**
  - **Type I error**
    - The null hypothesis is rejected when it should be accepted
  - **Type II error**
    - The null hypothesis is not rejected when it is not true

Decision \ Reality	H_0 true	H_1 true
H_0 accepted		
H_0 rejected		

Probabilities of Type I and Type II Errors(2)

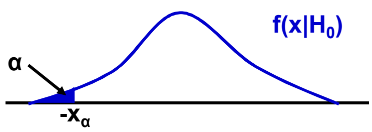
- **Properties**
  - **Type I error depends on probability α**
  - **Type II error depends on α, the sample size, the true value of parameters**

Type of Tests(1)

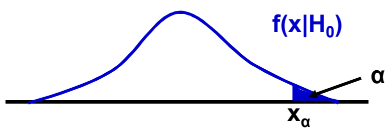
- **Null Hypothesis**

  H_0: $\mu_1 = \mu_2$
- **One-tailed Tests:**

  H_1: $\mu_1 < \mu_2$ (lower tail test) **Rejection Region: $X_{test} \leq -x_\alpha$**



  H_1: $\mu_1 > \mu_2$ (upper tail test) **Rejection Region: $X_{test} \geq -x_\alpha$**



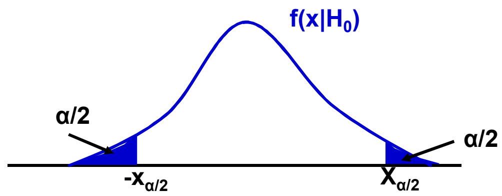
Type of Tests(2)

- **Two-tailed Test:**

H_1: $\mu_1 \neq \mu_2$

Rejection Region: $|X_{test}| \geq -x_\alpha$

Equivalently, if $X_{test} \geq x_{\alpha/2}$ or $X_{test} \leq -x_{\alpha/2}$



Two-tailed Example (1)

- **Hypothesis**
  - ***Null hypothesis H_0*: $\mu = \mu_0$**
  - **Alternative hypothesis H_1: $\mu \neq \mu_0$**
- **Test Statistic**
- **Standardized Test Statistic**

Two-tailed Example (2)

- **Type II error for a given α**
  - **The probability βof a Type II error is dependent on α, *n*, and *c/σ***

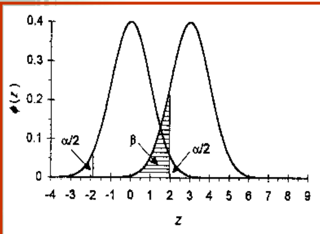


FIGURE 5.4.1
Significance test on the sample mean. The normal (0,1) pdf on the left is the assumed model in standardized form. Suppose the normal (3,1) pdf on the right represents the true model. Then the Type I error is α and the Type II error is β. The test is applicable regardless of the sign of the shift in the mean. The magnitude of β changes with the absolute magnitude of the shift; α is invariant here.

	H_0 true	H_1 true
Accept H_0	$1-\alpha$	β
Accept H_1	α	$1-\beta$

Two-tailed Example (3)

- **Characteristic Curves**
  - **For the same sample size *n*, β decrease as *c/σ* increases**
  - **β decreases as the sample size *n* increases**

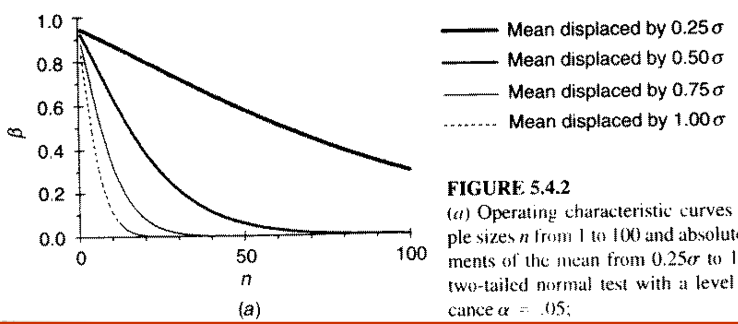
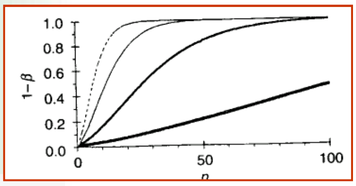


FIGURE 5.4.2
(*a*) Operating characteristic curves with sample sizes n from 1 to 100 and absolute displacements of the mean from 0.25σ to 1.00σ for a two-tailed normal test with a level of significance $\alpha = .05$;

Power Function

- Definition: Probability we do the right thing
- Properties
  - The complement of β
  - Probability of rejecting the null hypothesis when it is not true
- Power Function Example
- Power Curve Example



	H_0 true	H_1 true
Accept H_0	$1\text{-}\alpha$	β
Accept H_1	α	$1\text{-}\beta$

One-tailed Example 1: Testing *n*=1 beam (1)

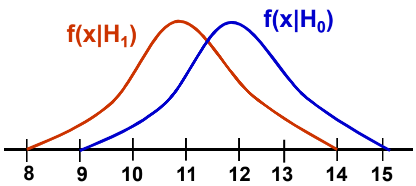
- Do we have the premium beams, or the regular beams?
  - Testing *n*=1 beam

    Premium: X ~ N[12, 1^2]

    Regular: X ~ N[11, 1^2]

    State #1 – H_0: μ= 12 (σ=1)

    State #2 – H_1: μ= 11 (σ=1)



One-tailed Example 1: Testing *n*=1 beam (2)

- **Need a Decision Rule – Once a week test a beam:**

Accept H_0: μ= 12 if $X > c_X$

Accept H_1: μ= 11 if $X \le c_X$

c= critical x-value for test

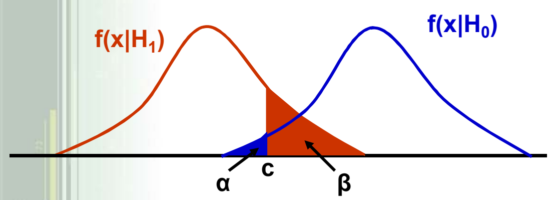
$X \le c$ is *rejection region* for H_0

- **Type I error**

One-tailed Example 1: Testing *n*=1 beam (3)

- **Type II error**

β= P[Accept H_0 | H_0 false]

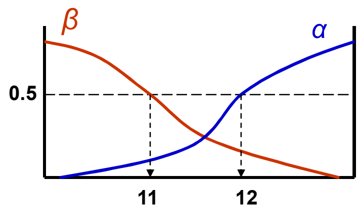


	H_0 true	H_1 true
Accept H_0	1-α	β
Accept H_1	α	1-β

One-tailed Example 1: Testing n=1 beam (4)

- **The Trade-off**

c_X	α	β
9	0.0013	0.98
10	0.023	0.84
11	0.16	0.50
12	0.50	0.16
13	0.84	0.023



Special values

10.72	0.10	0.76
9.67	0.01	0.91

One-tailed Example 2: Testing n=4 beam (1)

- **Now, what if we select n=4 beams to test?**
  Do we have the premium beams, or the regular beams?
  Use sample average to construct a decision procedure

  $X \sim N \rightarrow \bar{X} \sim N$

  $E[\bar{X}] = E[X]$; $Var[\bar{X}] = \sigma^2/n$

  ***Premium*: $\bar{X} \sim N[12, (1/4)]$**
  ***Regular*: $\bar{X} \sim N[11, (1/4)]$**

One-tailed Example 2: Testing *n*=4 beam (2)

- **New Trade-offs**

c_X	α	β
10	3×10^{-5}	0.98
10.84	0.01	0.63
11	0.023	0.50
11.36	0.10	0.24
12	0.50	0.023

Testing The Difference Between Two Means Using Known Variances

- **Assumptions**
  - X_1, X_2 ~ normal
  - σ_1 and σ_2 are known
- **Hypothesis**

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

- **Test Statistic**

where $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$

Testing The Difference Between Two Means When The Variances are Unknown But Equal

- **Assumptions**
  - **X_1, X_2 ~ normal**
  - **$\sigma_1 = \sigma_2$**
  - **σ_1 and σ_2 are unknown**
- **Hypothesis**

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

- **Test Statistic**

where $$\hat{S}_p = \frac{(n_1 - 1)\hat{S}_1^2 + (n_2 - 1)\hat{S}_2^2}{n_1 + n_2 - 2}$$

Testing The Difference Between Two Means When The Variances are Unknown and Unequal

- **Behrens-Fisher Problem**
- **Assumptions**
  - **X_1~ N(μ_1, σ_1) and X_2 ~ N(μ_2, σ_2)**
  - **$\sigma_1 \neq \sigma_2$**
  - **σ_1 and σ_2 are unknown**
- **Hypothesis**

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

- **Test Statistic**

Degree of freedom: $$\nu = \frac{[(\hat{S}_1^2 / n_1) + (\hat{S}_2^2 / n_2)]^2}{[(\hat{S}_1^2 / n_1)^2 / (n_1 - 1)] + [(\hat{S}_2^2 / n_2)^2 / (n_2 - 1)]}$$

Testing The Difference Between Two Means Example

Example 5.18. Change in the mean annual maximum flow with unknown and unequal variances. We return to the data of Example 5.15. The two main parameters are μ_1 and μ_2.

Null hypothesis H_0: $\mu_1 = \mu_2$.
Alternate hypothesis H_1: $\mu_1 > \mu_2$.
Level of significance $\alpha = .01$.

Calculations: We use a one-tailed test. From the data of Example 5.15,

$$\bar{x}_1 = 1569,\ \hat{s}_1 = 372,\ x_2 = 3653,\ \hat{s}_2 = 1563\ \mathrm{m^3/s} \qquad n_1 = n_2 = 12.$$

Hence from Eq. (5.4.11),

$$\nu = \frac{[(372^2/12) + (1563^2/12)]^2}{[(372^2/12)^2/11] + [(1563^2/12)^2/11]} \approx 12,$$

$$T = \frac{(3653 - 1569)}{\sqrt{(372^2/12) + (1563^2/12)}} = 4.49.$$

From Table C.2, $t_{12,0.01} = 2.681$.
Decision: The null hypothesis is rejected.

For others, see Table 5.4.1

Goodness-of-Fit Tests(1)

- **Chi-Squared Goodness-of-Fit Test**
  - **Main steps**
    - The ranking of a sample of data
    - Division into a number of classes depending on the magnitudes and the range
    - The fitting of a probability distribution
  - **Test Statistic**

where O_i: observed frequency, E_i: expected frequency
l: the number of class, k: the number of parameters

Chi-Square Example(1)

Example 8.6. As an example of using the chi-square test, consider the data of table 2.2 and test the hypothesis that the data are from a normal distribution. The observed and expected numbers in each class interval are obtained by multiplying the relative frequency by 66 which is the number of observations. Table 8.2 shows the calculation χ_c^2. The degrees of freedom is k-3 or 7 since two parameters (μ_X and σ_X^2) were estimated for the normal distribution. Comparing χ_c^2 of 10.31 with $\chi^2_{0.90,7}$ = 12.02, it is concluded that the normal distribution adequately describes the data for α = 0.10. If χ_c^2 had exceeded $\chi^2_{1-\alpha,k-p-1}$, the hypothesis that the normal distribution describes the data would be rejected.

Table 8.2. Chi-square test on Kentucky River data.

Class Mark	Observed Number	Expected Number	O-E	$\frac{(O-E)^2}{E}$
25,000	2	1.65	0.35	0.0742
35,000	3	3.76	-0.76	0.1536
45,000	10	7.07	2.93	1.2142
55,000	9	10.49	-1.49	0.2116
65,000	11	12.41	-1.41	0.1602
75,000	10	11.75	-1.75	0.2606
85,000	12	8.44	3.56	1.5016
95,000	6	5.35	0.65	0.0789
105,000	0	2.57	-2.57	2.5700
115,000	3	0.99	2.01	4.0809
Total	66	64.48	+1.52	10.3058

Chi-Square Example(2)

Table 8.3. Chi-square test on Kentucky River data (modified).

Class Mark	Observed Number	Expected Number	O-E	$\frac{(O-E)^2}{E}$
25,000	5	5.41	-0.41	0.0311
35,000				
45,000	10	7.07	2.93	1.2142
55,000	9	10.49	-1.49	0.2116
65,000	11	12.41	-1.41	0.1602
75,000	10	11.75	-1.75	0.2606
85,000	12	8.44	3.56	1.5016
95,000	6	5.35	0.65	0.0789
105,000				
115,000	3	3.56	-0.56	0.0881
Total	66	64.48	+1.52	3.5463

Chi-Square Example(3)

Table 8.4. Chi-square test based on equal expected numbers per class interval.

Class Number	Boundaries lower	Boundaries upper	Observed Number	Expected Number	$\frac{(O-E)^2}{E}$
1	$-\infty$	40,620	6	6.6	0.055
2	40,620	49,860	9	6.6	0.873
3	49,860	56,580	7	6.6	0.024
4	56,580	62,250	4	6.6	1.024
5	62,250	67,500	6	6.6	0.055
6	67,500	72,750	8	6.6	0.300
7	72,750	78,420	4	6.6	1.024
8	78,420	85,140	9	6.6	0.873
9	85,140	94,380	8	6.6	0.300
10	94,380	∞	5	6.6	0.388
		Totals	66	66.0	4.913

Goodness-of-Fit Tests(2)

- **Kolmogorov-Smirnov Goodness-of-Fit Test**
  - **Main steps**

    - A completely specified theoretical continuous cdf: $F_0(x)$

    - The empirical or sample distribution function: $F_n(x)$

    - The test statistics:

    - Reject if $D > D_{n,\alpha}$ (Table C.7)

Kolmogorov-Smirnov Example(1)

Example 5.29. Testing goodness-of-fit of timber strengths. In Table E.1.1, modulus of rupture data from 50 × 150 mm Swedish redwood and whitewood timber are given in newtons per square millimeter. Let us suppose that the first 100 items were delivered by one supplier, and a second lot of 64 items came from another batch. (We ignore the zero item at the end.) We assume that the distributions of the two lots are identical, but we will examine this in the next example. Because of the positive skewness shown in Table 1.2.3, the distribution can be one of several types, such as the gamma described in Chapter 4. The Weibull seems to be an ideal candidate, however, because it was originally devised to model material strengths and similar effects and has been used for such purposes for over 65 years. We apply, in the first instance, the two-parameter Weibull distribution with cdf, as presented before through Eq. (4.2.16),

$$F_X(x) = 1 - \exp\left[-\left(\frac{x}{\lambda}\right)^{\beta}\right].$$

Following the least squares estimation procedure of Example 4.25, we estimate the following parameters from the first 100 items of data:

$$\hat{\beta} = 5.39 \quad \text{and} \quad \hat{\lambda} = 42.55.$$

Kolmogorov-Smirnov Example(2)

We use these estimates to model the empirical distribution of the next 64 items and hence apply the Kolmogorov-Smirnov goodness-of-fit test.

Null hypothesis H_0: The random variable (representing the modulus of rupture of 50 × 150 mm Swedish redwood and whitewood timber) has a Weibull distribution as specified earlier.

Alternate hypothesis H_1: The random variable has a different distribution.

Level of significance $\alpha = .05$.

Calculations: The critical region D_n, as defined by Eq. (5.6.4a), is

$$> D_{n,\alpha} = 1.3581/\sqrt{n} = 1.3581/\sqrt{64} = 0.17.$$

By setting this value above and below the sample cdf, confidence limits can be drawn as shown in Fig. 5.6.1. It is also shown that $d_n = 0.1008$, which is less than the critical value.

Decision: The null hypothesis is not rejected.

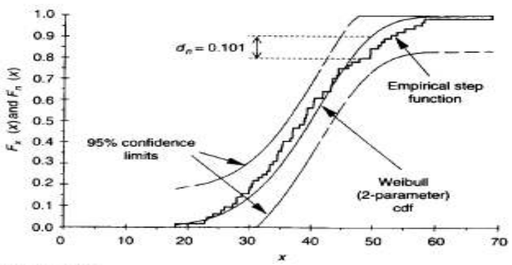


FIGURE 5.6.1
Kolmogorov-Smirnov one-sample goodness-of-fit test.

Goodness-of-Fit Tests(3)

- **PPCC Test**
  - **Test Statistics**

where, x_i: ranked observation, w_i: fitted qunatile(=$G^{-1}(1-q_i)$)

r: correlation coefficient, q_i(=p_i): plotting position

G(x): proposed cdf for the events

Goodness-of-Fit Tests(4)

- **Lower Critical Values for PPCC Test**

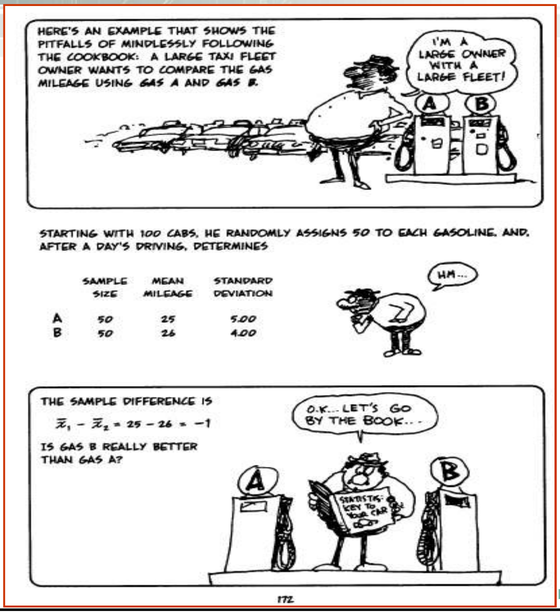
TABLE 18.3.3 Lower Critical Values of the Probability Plot Correlation Test Statistic for the Normal Distribution Using $p_i = (i - 3/8)/(n + 1/4)$

	Significance level		
n	0.10	0.05	0.01
10	0.9347	0.9180	0.8804
15	0.9506	0.9383	0.9110
20	0.9600	0.9503	0.9290
30	0.9707	0.9639	0.9490
40	0.9767	0.9715	0.9597
50	0.9807	0.9764	0.9664
60	0.9835	0.9799	0.9710
75	0.9865	0.9835	0.9757
100	0.9893	0.9870	0.9812
300	0.99602	0.99525	0.99354
1000	0.99854	0.99824	0.99755

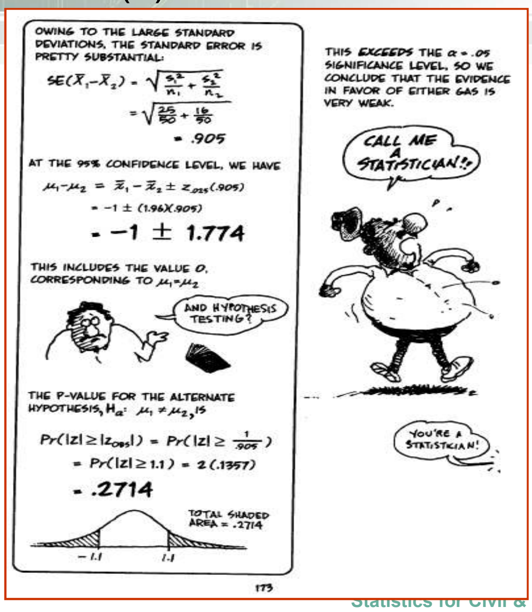
TABLE 18.3.4 Lower Critical Values of the Probability Plot Correlation Test Statistic for the Gumbel and Two-Parameter Weibull Distributions Using $p_i = (i - 0.44)/(n + 0.12)$

	Significance level		
n	0.10	0.05	0.01
10	0.9260	0.9084	0.8630
20	0.9517	0.9390	0.9060
30	0.9622	0.9526	0.9191
40	0.9689	0.9594	0.9286
50	0.9729	0.9646	0.9389
60	0.9760	0.9685	0.9467
70	0.9787	0.9720	0.9506
80	0.9804	0.9747	0.9525
100	0.9831	0.9779	0.9596
300	0.9925	0.9902	0.9819
1000	0.99708	0.99622	0.99334

Cartoon(1)



Cartoon(2)



Cartoon(3)

PAIRED COMPARISONS
a better way to compare gasolines

THE TAXI OWNER FOLLOWED THE COOKBOOK EXACTLY. HIS SAMPLES WERE RANDOM, AND HIS SAMPLE SIZE WAS LARGE ENOUGH. HE JUST FAILED TO *THINK* WHEN NECESSARY!

ALTHOUGH GAS B APPEARS TO BE SLIGHTLY BETTER THAN GAS A, THE CONFIDENCE INTERVAL WAS WIDE BECAUSE OF THE LARGE STANDARD DEVIATIONS—I.E., THE *MILEAGES VARIED WIDELY FROM ONE CAB TO THE NEXT.* WHY SUCH HIGH VARIABILITY? BECAUSE CABS—AND CABBIES—HAVE DIFFERENT *PERSONALITIES!*

174

Cartoon(4)

A FAR BETTER WAY TO DO THIS STUDY IS TO ASSIGN GAS A AND GAS B TO THE *SAME CAB ON DIFFERENT DAYS.*

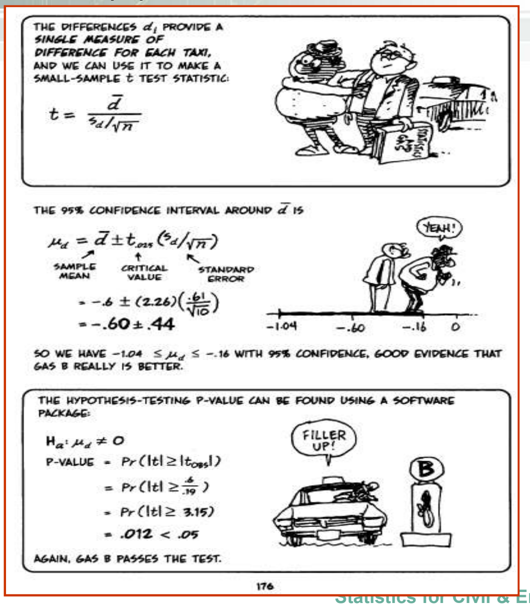
WE STILL RANDOMIZE THE TREATMENT BY FLIPPING A COIN TO DECIDE WHETHER TO USE GAS A ON TUESDAY OR WEDNESDAY. WE CAN ALSO CUT THE EXPERIMENT DOWN TO 10 CABS, SAVING THE OWNER A LOT OF TIME AND MONEY!

CAB	GAS A	GAS B	DIFFERENCE
1	27.01	26.95	0.06
2	20.00	20.44	−0.44
3	23.41	25.05	−1.64
4	25.22	26.32	−1.10
5	30.11	29.56	0.55
6	25.55	26.60	−1.05
7	22.23	22.93	−0.70
8	19.78	20.23	−0.45
9	33.45	33.95	−0.50
10	25.22	26.01	−0.79
MEAN	25.20	25.80	−0.60
STANDARD DEVIATION	4.27	4.10	0.61

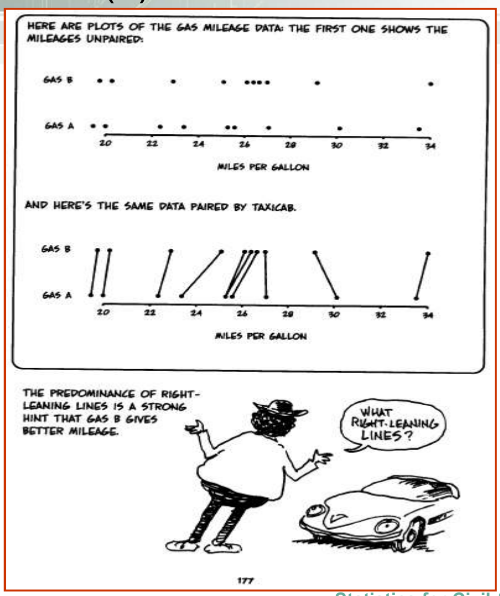
NOTE THAT THE MEANS AND STANDARD DEVIATIONS OF GAS A AND GAS B ARE ABOUT THE SAME. THAT'S TO BE EXPECTED, SINCE THEY HAVE THE SAME SOURCE OF VARIABILITY AS IN THE UNPAIRED EXPERIMENT. BUT NOW THE *DIFFERENCE COLUMN* HAS A VERY *SMALL* STANDARD DEVIATION. THE DIFFERENCE COLUMN, BY COMPARING GAS PERFORMANCE WITHIN A SINGLE CAR, ELIMINATES VARIABILITY BETWEEN TAXIS.

175

Cartoon(5)



Cartoon(6)



Cartoon(7)

A PAIRED COMPARISON EXPERIMENT IS ONE OF THE MOST EFFECTIVE WAYS TO REDUCE NATURAL VARIABILITY WHILE COMPARING TREATMENTS. FOR EXAMPLE, IN COMPARING HAND CREAMS, THE TWO BRANDS ARE RANDOMLY ASSIGNED TO EACH SUBJECT'S RIGHT OR LEFT HANDS. THIS ELIMINATES VARIABILITY DUE TO SKIN DIFFERENCES.
A
B
A
B
A
B
OR, IN COMPARING TWO BREAKFAST CEREALS, EACH TASTER RATES BOTH CEREALS (IN RANDOM ORDER). THE PAIRED COMPARISON REMOVES THE NATURAL BIAS OF THE TASTER FOR OR AGAINST CEREAL IN GENERAL.
PHAW! WHAT EVER HAPPENED TO BACON AND EGGS?
NEW AGE GRANOLA
OLD CORN COBS
178