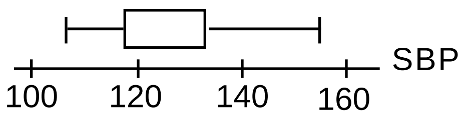
1[5] In a class of 100 students, the grades on a test are summarized in the following frequency table.

Circle the interval which contains the median grade:

Grade	Frequency	Cumul.
91 -100	11	
81 - 90	31	
71- 80	42	58 (>50%)
61 - 70	16	16 (<50%)

cross 50% in the 71-80 interval

2[5] Consider this box plot of the systolic blood pressures (SBP) of 39 adult males.



Based on this box plot...

(i) the interquartile range is approximately (circle one):

50 **15** 125 70

(ii) which of these statements is true?

(circle one)

a The minimum blood pressure is less than 100.

b **The max blood pressure is about 155.**

c There is an outlier with a value of about 170.

d The median blood pressure is about 115.

3[5] A sample was taken of the salaries of 20 employees of a large company. The following are the salaries (in thousands of dollars) for this year. For convenience, the data are ordered.

28 31 34 35 37 41 42 42 42 47 47 51 52 52 60 61 67 72 75 80

(i) Suppose each employee in the company receives a $3,000 raise for next year (each employee's salary is increased by $3,000). The standard deviation of the salaries for the employees will (circle one) :

a **be unchanged.**

SD(Y+c) = SD(Y)

b increase by $3,000.

c be multiplied by $3,000.

d increase by $\sqrt{3,000}$.

(ii) Suppose each employee in the company receives a 4% raise for next year. The standard deviation of the salaries for the employees will (circle one) :

a be unchanged.

b increase by 16%

c **increase by 4%**

SD(c Y) = c SD(Y)

d increase by 2%

4[5] Suppose that the probability that HIV will be passed from an infected person to an uninfected person during a single sexual contact is 0.01. Suppose that there are 50 such contacts.

Show how to calculate (or obtain via software) the probability that HIV will be passed on in at least one of the 50 contacts.

1 - Prob(escape 50 times) = 1 - $(0.99)^{50}$ [like Russian Roulette but with 99 blank chambers and 1 loaded chamber !)

Can also use
1 - BinomialProb(x=0, p =0.01, n = 50, don't cumulate)

5[8] The following are data collected by Statistics Canada at the 1996 census:

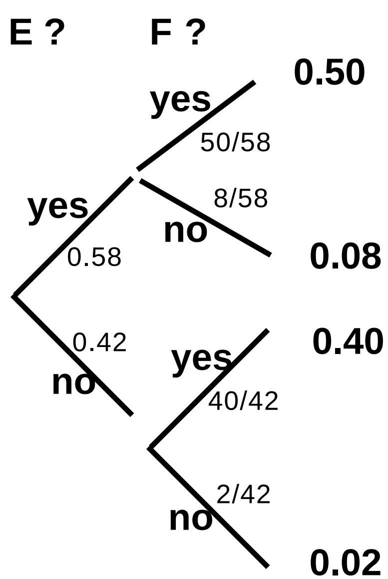
Montreal Metropolitan Population by knowledge of official language

English only	French only	Both English and French	Neither English nor French
8%	40%	50%	2%

Consider a person randomly chosen from this population. Let E be the variable denoting knowledge(Yes/No) of English, and F knowledge (Yes/No) of French.

a) *Display the joint distribution of E and F as a 2 x 2 table or as a probability tree.*

		French		
		Yes	*No*	
English?	*Yes*	*0.50*	*0.08*	*0.58*
	No	*0.40*	*0.02*	*0.42*
		0.90	*0.10*	*1.00*



b) *What is the probability that :*

(i) a person speaks French

90%

(i) a person who knows French also knows English

50/90

(ii) a person who knows English also knows French

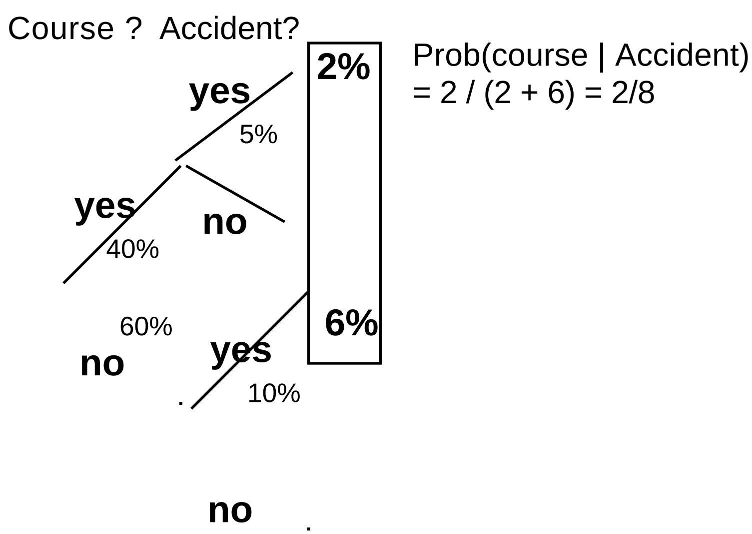
50/58

c) *Are E and F independent random variables? Why or why not?*

NO, since 0.90 x 0.58 does not equal 0.50

6[8] An auto insurance company notes whether drivers under 25 years old have had a driver's education course. Some 40% of its policyholders under 25 have had a driver's education course and 5% of this subset have an accident in a one-year period. Of those under 25 who have not had a driver's education course, 10% have an accident within a one-year period.

A 20-year-old takes out a policy with this company and within one year has an accident. What is the probability that the person did not have a driver's education course? [a probability tree may help]



7[5] Suppose that 10% of McGill students are left-handed.

Find the probability that in a class of 20 students, at least 2 will be left-handed. State any assumptions made.

Binomial with n=20, p = 0.1.

Prob (≥ 2)= 1 - (Prob[0 or 1]) = 1 - (Prob[0] + Prob[1])

= 1 - (0.1216 + 0.2702) = 0.6082

8[10] To allow for the 10% (on average) of people who make reservations but do not show up for their flights, an airline company "over-books" i.e., it takes reservations for up to 145 seats when it only has space on the plane for 140 passengers.

a) *If the airline takes 145 reservations, what is the probability that it will not have a seat for every passenger who shows up? Actual calculations are not required, but explain the steps in sufficient detail that your research assistant could complete them.*

Binomial with n = 145, Prob[show] = 0.90

Prob[not enough] = Prob[141 or 142 or .. or 145]

= P[141] + ... + Prob[145]

= 1 - BINOMDIST[x=140,n=145,p=0.9, cum. = TRUE)

b) *Would the normal distribution be appropriate here? Why or why not?*

YES, since the distribution is 'well in' from the 2 corners

Expected No. of Shows = 145 x 0.9 = 130.5

Expected No. of No Shows = 145 x 0.10 = 14.5

c) *Suppose that some of the 145 individuals are related. Does this affect the appropriateness of the probability model you base your calculations on? Why/why not?*

Would not be independent.. persons in same party would probably act in same way.

9[4] An airplane has a front and a rear door which are both opened to allow passengers to exit when the plane lands. The plane has 100 passengers seated.

(i) The number of passengers exiting through the front door should have (circle one):

a a binomial distribution with mean 50.

b a binomial distribution with 100 trials but success probability not equal to 0.5.

c a normal distribution with a standard deviation of 5.

d **none of the above.**

(ii) justify your choice

Probabilities vary from near 1 at front to near 0 at back. May well be 1 for the 35 nearest the front, and 0 for say the 35 nearest the rear, so only uncertainty is in the 30 in the middle.

[a bit like Yogi Berra.. asked how many of the 162 games his baseball team would win: he replied: I can tell you for sure that we will win 50 and lose 50.. its the other 62 that I'm not sure about..."]

Also, independence and unequal probabilities are different concepts

10[8] Suppose we want a 95% confidence interval for the average amount spent on books by students in their first year at university. The interval is to have a margin of error of $5.

(i) If the 'best estimate' is that the amounts spent have a standard deviation σ = $50, then the number of observations required (in a SRS) is closest to:
(circle one)

a 196

b 250

c **400**

d 475

$n = (1.96)^2 s^2 / m^2 = \text{(approx.) } 2^2 \, 50^2 / 5^2 = 400$

(ii) *The amounts spent by individuals are unlikely to have a normal distribution. Is this critical to your calculation? Explain.*

No, since with this size sample, the Central Limit theorem will ensure that the possible sample means will have a Gaussian distribution (remember the very skewed insurance earnings data (individual policies) but the Gaussian behaviour of ybar when n = 400.

mean is a statistic and so its distribution has a much more regular shape than individual values do

10...continued

(iii) The university administrator, who fancies himself as a statistician, says that we should use a multiple from the "t" distribution rather than the z distribution in the sample size formula.

Technically, he is correct. Explain why.

have to estimate SD from s in the actual sample [won't use guesstimated "σ "= 50] use t whenever estimate σ.

Practically, in this case, it doesn't matter very much. Explain -- in a sentence or two -- why.

t --> z as df -> large (≈ 400 here)

11[8] The nicotine content in cigarettes of a certain brand is normally distributed with mean (in milligrams) μ and standard deviation $\sigma = 0.1$. The brand advertises that the mean nicotine content of its cigarettes is 1.5, but measurements on a random sample of 100 cigarettes of this brand gave a mean of $\bar{x} = 1.53$. Is this evidence that the mean nicotine content is actually higher than advertised?

(i) To answer this, test the hypotheses

H_0: $\mu = 1.5$, H_a: $\mu > 1.5$

at the 5% significance level. *SHOW YOUR WORK*

1 sided
α **= 0.05**
σ **given so can use z test**

z = (1.53 – 1.50) / { 0.1/sqrt[100] } = 3

prob (Z > 3) = .0013 which is < α.

(ii) You conclude (circle one) :

a **that H_0 should be rejected.**
b that H_0 should not be rejected.
c that H_a should be rejected.
d there is a 5% chance that the null hypothesis is true. **anyone who says this gets a note to "watch your language"**

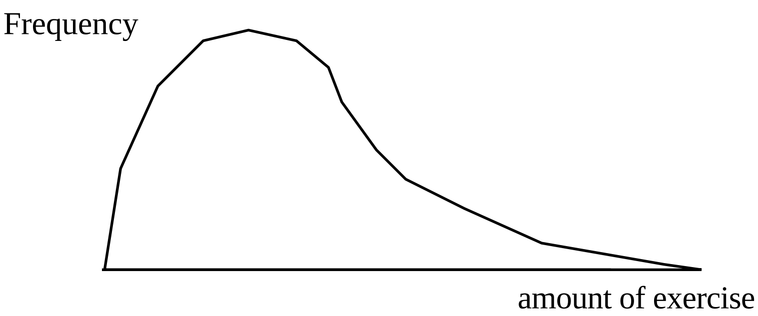
12 Refer to the article "Oral contraceptive use and bone mineral density in premenopausal women" As the authors did, concentrate on Never vs. Ever (Table 1, first two columns)

a[5] The reported SD for alcohol consumption is larger than the mean.

(i) *Can this be, or is it a mistake? Explain, using a diagram to show what the distribution looks like.*

Yes it can be; not necessarily a mistake (probably) a long R tail; maybe some who don't exercise at all

Could not be Gaussian, since mean minus 1SD or minus 2SD would put some into the negative exercise territory!



(ii) What other summary measures might have been more informative for describing the shapes of the distributions of alcohol and cigarette consumption?

% who don't exercise at all
5 number summary

Q1 Q2 Q3
median
——————————
IQR

etc

b[5] *Reconstruct, from the data given, the 95% CI of (–0.06 to 0.00) that accompanies the point estimate of –0.03 for the lumbar BMD.*

1.03 – 1.06 ± 1.96 sqrt[0.13^2 / 64 + 0.13^2 / 429]

= -0.03 ± 0.034 ± 0.034

(-0.064 to +0.004) author rounds to –0.06 and +0.00

(1.96, effectively z, since large df)

If use df < ∞ e.g. min. (63, 428) = 63 df,
multiple is 2.00 approximately, so CI slightly wider

c[5] The authors do not report any P-values, but in at least two places (3 lines from bottom of first column of page 1025, legend to Figure 1) they do use the concept of statistical significance.

Without doing any calculations, deduce directly from what is given in the table the (approximate) 2-sided P-value associated with the 1.03 versus 1.06 for the lumbar BMD. Explain your reasoning.

95% CI just touches 0, so P = 0.05 (2 sided)

d[5] If you were calculating a P-value from the 1.03 versus 1.06, you would have had to decide between a "pooled variance" test or a "separate sample" t-statistic.

Give two reasons why -- in this instance -- the P-value would have been just about the same either way.

two s's are virtually same (to 2dp), so adjustment to df has minimal impact

with large df, e.g., even at 63 (min) t -> z

e[5] The authors do not explain what the "error bars" on the top of the rectangles in Fig. 1 represent. By eye, I estimate that the lengths of these bars for the two "Spine" BMD's are of the order of 0.08 g/cm^2.

What could they be? Or if you can't be sure, say what they could not be. Explain your reasoning.

I don't think they are SEM's (both same size, even with very different n's.

Know s = 0.13 (or less if adjusted, i.e. if effect of age etc. removed)

Try it out...

SEM n = 64 $$\frac{0.13}{\sqrt{64}} = \frac{0.13}{8} \approx .016$$

So even 2 SEM's would only be 0.032 and if n = 427, SEM would be tiny.

(must be some type of SD for individuals)

f5] Suppose you were told that the Lumbar BMD of one of the 524 women was -- when rounded -- 1.35g/cm^2.
What is the probability that the woman was in the "Ever" group? (use the information below)
8.5 people out of 524 are in this BMD category; 6.7/8.5 are from the "ever" and 1.8 from "never"
so P (ever | BMD) = 6.7 / (6.7 + 1.5) = 6.7/8.5 ≈ 78.8%

[a little lower if use rounded rather than unrounded values]

Lumbar BMD categories

Prior Prob.	OC												
		0.65	0.75	0.85	0.95	1.05	1.15	1.25	**1.35**	1.45	1.55	1.65	(1)
0.87	**Ever**	0.00	0.03	0.12	0.25	0.30	0.20	0.07	**0.01**	0.00	0.00	0.00	(2E)
# 454													
		1.9	13.7	53.4	115	138	91.0	33.3	**6.7**	0.8	0.0	0.0	(3E)
									???				
									???				
		0.1	1.3	5.8	15.0	21.4	16.9	7.4	**1.8**	0.2	0.0	0.0	(3N)
# 70													
0.13	**Never**	0.00	0.02	0.08	0.21	0.31	0.24	0.11	**0.03**	0.00	0.00	0.00	(2N)
		2.1	14.9	59.2	130	159	108	40.7	**8.5**	1.0	0.1	0.0	(T)

Rows (1) : categories of Lumbar BMD (i.e., BMD rounded to end in a 5)
(2E) : Prob(BMD category | Ever) (Σ = 1)
(2N) : Prob(BMD category | Never) (Σ = 1)
(3E) : 454 × (2E) (Σ = 454)
(3N) : 70 × (2N) (Σ = 70)
(T) : totals of 3E and 3N (for each Lumbar BMD category)

13 Refer to the article "The effect of working serial night shifts on the cognitive functioning of emergency physicians"

a[12] "The mean day-shift KAIT score was 119.1 (SD=7.7), and the mean night-shift KAIT score was 107.2 (SD=10.2). This difference was significant (mean difference=11.9; 95% confidence interval 7.0 to 16.8; *P*<.001), with the day-shift scores being statistically higher than the night-shift scores" (Abstract; but see also more complete summaries in Table 1)

(i) *Reconstruct the 95% CI 7.0 to 16.8 from the summaries given.*

<u>16 pairs</u>

11.9 ± $t_{15.95}$ 9.2/√16
± 2.131 (2.3)
± 4.9

11.9 – 4.9 = 7
11.9 + 4.9 = 16.8

(ii) State the null and alternative hypotheses tested and verify that "P< .001"

$\mu_{diff} = 0$ vs. $\mu_{diff} \neq 0$

t = (11.9 – 0)/{9.2/√16} = 5.17

15 df	0.025	.02	.01	.005	.0025	.001	.0005	
1 tail	2.131	2.249	2.602	2.947	3.286	3.733	4.073	| 5.17

so: p < 0.0005 1-sided
p < 0.001 2-sided

(iii) Why, in the last row of the Table, doesn't $\sqrt{7.7^2 + 10.2^2}$ *equal 9.2 ?*

Because 2 variables are not independent, they are <u>positively</u> correlated;

so var[diff] < sum of variances

13 ... continued

b[12] "Residents in group B, who were tested first after working night shifts, had a larger difference between their 2 scores than residents in group A, who were tested first on the day shift (night first: mean difference=17.1 [SD=8.6]; day first: mean difference= 6.6 [SD=6.7]; *P*=.017). On the basis of these scores, the order of testing with the KAIT (night first or day first) did make a difference" [Bottom of page 153 and top of page 154]

(i) Reconstruct the P-value (0.017) from the summaries given.

$t = (17.1 – 6.6) / \text{sqrt}[8.6^2/ 8 + 6.7^2/ 8] = 10.5/3.85 = 2.72$

	025	.02	.01	.005	
using 14 df 1-sided p	2.14	2.26	2.62 |	2.98	close to 0.008
using 7 df 1-sided p	2.36	2.52 |	3.00	3.50	close to 0.015

2.72 close to 2.62 in 14 df table;

p-value 1-sided approximately 008, 2-sided 016

(I think they used 14 df, and 2-sided)

If they used separate samples, df would between 7 & 14 (somewhere)

If used pooled covariances, you are suing an s^2 that is 1/2 way between 8.6^2 and 6.7^2 i.e., between 73.96 and 44.89, i.e. 59.43 and √59 = 7.7

10.5/{7.7 sqrt[1/8 + 1/8] = 10.5 / 3.85 is the same

= careful! Not always easy to do. Can’t get people “back” to initial state.

(ii) Explain in words -- to a resident who is working the day shift -- what the P-value of 0.017 is (after the night shift, don't even try!).

If there were no difference between two ways of measuring the difference, then the chance that we would observe a <u>discrepancy</u> of 10.5 <u>or more</u> is only 0.017.

(iii) Why did the order of testing make a difference? What is the lesson for investigators who are attracted to the crossover design?

learning effect

careful! Not always easy to do. Can’t get people “back” to initial state.

They had the possibility of removing the learning effect, by subtracting the almost 8 points from the day scores in Fig 2, and the same amount from the night scores in Fig 1. As it stands now, the 16 differences contain this 'noise' -- it is one of the reasons the SD of the 16 differences is so high -- 9.2.

Fortunately, the signal was still strong enough t shine through, despite the noise. A smaller signal might not have made it.

As for 'bias' or 'confounding', the authors took care of this with the counterbalanced design.. 8 done in each order. Had they done 12 in one order and only in the other order, then the average difference would contain some learning effect and some day vs. night effect.