1. Admissions to the emergency ward

$\sum f = 50$ $\sum fx = 120$ $\sum fx^2 = 394$; $120^2 / 50 = 288$

a modal x = 2; mean x = 120/50= 2.4 admissions

median x = 2 ... the middlemost of the 50 x's arranged in numerical order: both the 25th and 26th are 2.

00000111111112222222222222223333333333444444555 6

range of x = 0 to 6

Variance of x = s^2=(394 -288)/49 = 2.16; sd of x = s = 1.47

b If we addded the number of admissions for a 51st day, only the mean would change; mode and median would not. The variance, range, sd are not statistics of central tendency.

c If observations for 100 days rather than 50 days were contained in the dataset, you should expect the range to be larger because it can only get wider, and with 50 more, it probably would. However, one cannot predict whether the standard deviation would get larger or smaller. It does not vary systematically with n.

d The coefficient of variation =[sd/mean] x 100%.

e To compare the variability in the number of admissions to this ward with that in another hospital that handles more emergencies, the coefficient of variation would help neutralize the fact that the SD might be proportionately larger.

2. The probability that all three cities have fog on the same day in January? is the product of the 3 given probabilities. i.e.

   $P(ABC) = P(A) \cdot P(B|A) \cdot P(C|A \text{ and } B)$

   Note that the events do not have to be independent. In fact they are not, and that is why we resort to conditional probabilities.

   Incidentally, one person took the "same day in January" to mean "either Jan 1 or Jan 2 or ... or Jan 31" rather than "one designated day". With this interpretation, the probability is a lot higher.

3. The standard normal random variable Z.

   a $\text{Prob}(Z \leq .44) = 0.67$
   b $\text{Prob}(-.44 \leq Z \leq 0.44) = 0.33$
   c The value such that $\text{Prob}(Z \leq \text{value}) = 0.25$ is -0.67.

4. How there can be a sampling distribution of $\bar{x}$ when in practice only a single sample is selected?

   One could imagine the $\bar{x}$'s from all the possible random samples of size n. It would have a distribution around μ and with a spread given by $SE(\bar{x})$. It is this "predictable uncertainty" that allows us to predict how far (at most) our one sample $\bar{x}$ is likely to be from μ.

5. The Central Limit Theorem is "Central" and important in inferential statistics because it allows us to make inferences from sample means (proportions etc) using Gaussian tables even when the individual data are not Gaussian distributed. It says that the sampling distribution of $\bar{x}$ is Gaussian even if the x's are not (if n large enough...)

6. With simple random sampling, what distribution does each of the following statistics follow if:

	$\bar{x}$	$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$	$\frac{\bar{x} - \mu}{s/\sqrt{n}}$
X is normally distributed and n is small?	$N(\mu,\sigma_X/\sqrt{n})$	N(0,1) or "Z"	t
X is normally distributed and n is large?	$N(\mu,\sigma_X/\sqrt{n})$	N(0,1)	N(0,1)
X is not normally distributed is small?	Can't say	Can't say	Can't say
X is not normally distributed and n is large?	$N(\mu,\sigma_X/\sqrt{n})$	N(0,1)	N(0,1)
	<---------- Central Limit Theorem------>		

N = Normal (Gaussian)

7. A research report (based on a random sample of hospitals) stated that the mean percentage reduction in the number of hospital beds over the past year was between 6.1% and 10.8% with a confidence coefficient of 95%. A interpreted this as meaning that 95% of hospitals had reductions in the number of beds between 6.1% and 10.8%. B interpreted this in the sense that if many random samples were taken, 95% of the samples would have mean reductions between 6.1% and 10.8%.

   Why are these interpretations incorrect?

   A: is talking about ± 2 SD's and about individual hospitals.

   B: The CI should have a 95% chance of including μ. Each of the samples B speaks about will have a mean and true enough 95% of them will be within ± 2SEM of μ but each CI will be in a different place (think of the diagram I showed in class of 45 CIs and how most of them included μ.

8 To halve the standard error of the mean, one must double √n i.e. quadruple n. SEM is proportional to $1/\sqrt{n}$ not to 1/n

9. A survey to estimate the proportion of 16 year old schoolgirls in Quebec that are protected against rubella examined 40 randomly selected girls from each of 25 randomly selected schools. The authors used the Binomial distribution

with n=1,000 to calculate the uncertainty of their estimated proportion.

What is wrong with this method of calculating the uncertainty?

We do not have 1000 independent observations. Rubella is an infectious disease and vaccination programmes are often carried out on a school by school basis.

The real uncertainty is much bigger because of the smaller n: we could easily choose clusters that have less than (or greater than) the average % protected.

10. What is wrong with a study protocol which states

H_0: $\bar{x} = 10.0$ H_{alt}: $\bar{x} \neq 10.0$

Statistical inference is about population parameters ie μ's.

11. Many journal editors and referees are suggesting that a confidence interval (CI) is often a useful adjunct (or even a replacement for) a statistical test of significance. Suppose that we use a t-test to decide whether two sample means are significantly different.

If the means ARE significantly different from each other, what more can the CI tell us?

How big the true difference is, whether it includes differences that are big enough to matter or whether the significance is an artifact of a large n and a small d (ie of no clinical significance)

If the means ARE NOT significantly different from each other, what more can the CI tell us?

If the study still rules in important differences or whether it is definitively negative (as it would if the CI were quite narrow around 0)