Stats Examples: Answers 3

1. Let X = length of random rod in meters, so $X \sim N(3.1, 0.15^2)$.

- $P(X > 3.42) = 1 - P(X < 3.42) = 1 - P(Z < [3.42 - 3.1]/0.15) = 1 - 0.984 \approx 0.016$ (from tables)
- $P(X < 3) = P(Z < [3-3.1]/0.15) = 1-P(Z < 2/3) = 1-[0.7454+0.67\times(0.7486-0.7454)] = 1-.74754 \approx 0.252$ [interpolating from table, which is not really required here]
- $P(3 < X < 3.2) = P(-2/3 < Z < 2/3) = P(Z < 2/3) - P(Z < -2/3) = P(Z < 2/3) - P(Z > 2/3) = P(Z < 2/3) - [1 - P(Z < 2/3)] = 2P(Z < 2/3) - 1 \approx 2 \times .74754 - 1 \approx 0.495$

$P(X < x) = 0.95 \implies Q = 0.95 \implies Z \approx 1.65 \implies x \sim 3.1 + 1.65 \times 0.15 \approx 3.35$. Note when roughly approximating an unknown distribution, there is no point quoting results to high precision since more than 2 significant figures is very unlikely to be accurate, so if the result is only needed approximately you don't need to interpolate (for if you want 3 or 4 significant figures you do).

2.

- The time for stages 1 and 2 to complete is the sum of two Normal variates, with $t_2 = X_1 + X_2 \sim N(50, 2 \times 3.8^2)$. Hence $P(t_2 > 45) = P(t_2 < [55 - 50]/5.38) = P(Z < 0.93) \approx 0.82$.
- $P(X_1 + X_2 + X_3 > t) = 0.08$, where $t = X_1 + X_2 + X_3$ is the completion time, with $t \sim N(75, 3 \times 3.8^2)$. So we want $P(Z < z) = 0.92 \implies Z = 1.4 \implies X_1 + X_2 + X_3 = 75 + 1.4 \times \sqrt{3}3.8 \approx 84$. So the penalty is about 84 hours.

3.

- By the central limit theorem, the sum of 50 marks is expected to be roughly $N(50\mu, 50\sigma^2)$, where $\sigma = 15$, and hence the average mark is $\bar{M} \sim N(\mu, 15^2/50)$. So the standard deviation is $15/\sqrt{50} \approx 2.1$.
- Since $\bar{M}$ is approximately normal, $\bar{M}_1 - \bar{M}_2 \sim N(\mu_1 - \mu_2, 2\sigma^2/50)$. If the groups have the same arrangements $\mu_1 = \mu_2$, hence $\bar{M}_1 - \bar{M}_2 \sim N(0, 9)$
- For a normal distribution 95% of the probability lies between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$, where for differences due only to random fluctuations $\sigma = \sqrt{9} = 3$ (from the previous answer). So we want $|\bar{M}_1 - \bar{M}_2| \gtrsim 1.96 \times 3 \approx 6$.

4. Let X = number of bits in error.

- $X \sim B(10^7, 5 \times 10^{-6}) \approx N(np, np(1-p)) \approx N(10^7 \times (5 \times 10^{-6}), 10^7 \times (5 \times 10^{-6}) \times (1 - 5 \times 10^{-6})) \approx N(50, 50)$.
- $P(45 \leq X \leq 55) \approx P(X < 55.5) - P(X < 44.5) = P(Z < [55.5 - 50]/\sqrt{50}) - [1 - P(Z < [55.5 - 50]/\sqrt{50})] = 2P(Z < 0.778) - 1 = 2 \times 0.78 - 1 \approx 0.56$
- Let Y = errors in chunk of 10^6 bits, then $Y \sim N(5, 5)$. Then $P(Y < 13) \approx P(Z < [12.5 - 5]/\sqrt{5}) = P(Z < 3.35) \approx 0.9996$. Hence the probability there being any failures is $1 - P(Z < 3.35)^{10} \approx 0.004$.

5.

i Normal should be quite accurate by the central limit theorem. Most likely failure is that the diameters are not independent, e.g. a calibration error in a machining tool could make them all come out too large in one batch until the error is corrected.

ii Binomial if they are independent (fail, not fail). Again likely to fail as may not be independent, e.g. due to failure of a part in the manufacturing system, or correlations between nearby items (e.g. oven temperature drifts over time, so batches of 3 items tend to have been made at similar different temperature, but when it gets too hot suddenly all the items fail for a while).

iii Nothing simple: some mixture of discrete (drinker, non-drinker) and a continuous distribution describing how much each drinker drinks.

iv If asteroid collisions of a given size are random in time (Poisson), the time till the first one should be given by an exponential distribution. Likely failures include correlations between asteroid arrivals with the position of the moon, time of year and phases of the other planets. The probability of a certain number of deaths is also expected to be correlated with population growth (and changes in distribution).

v The vote is an average of each poll sample, so by the central limit theorem should be approximately Normally distributed if the samples are all statistically the same. Likely failures include systematic differences in the sampling method used by different polling companies (e.g. different corrections for the distribution between the sexes and earnings; reliance or not on having a phone line rather than mobile), and also subtle differences in which question is asked.

vi If failures are independent, with probability p, the probability of having failure on the nth examination is $P(n) = (1-p)^{n-1}p$. (this is sometimes called the *geometric distribution*). As previously most likely to fail due to correlations, or p depending on n (e.g. machine parts wear out for large n).

6. The bus arrivals are a Poisson process with one on average every 15 minutes. The time until the first arrival t is then an exponential distribution, $f(t) = \nu e^{-\nu t}$, with $\nu = 4$/hour (see notes).

1 $\langle t \rangle = 1/\nu = 15$ minutes

2 By symmetry this is the same, $\langle t_{\rm prev} \rangle$ =15 minutes.

3 $\langle t_{\rm prev} + t \rangle = \langle t_{\rm prev} \rangle + \langle t \rangle = 30$ minutes.

4 The average time between buses is 15 minutes, so the average number of passengers the bus driver will report is 15. But the average time since the last bus when you catch one is 30 minutes, so the average number seen by passengers is 30. This is consistent because there are more people on the fuller buses. Equivalently, in the cases when the gap between buses is small, you are unlikely to happen to arrive in the short time between, you are most likely to arrive in one of the longer gaps.

For the interested or confused reader: In detail, say N_i is the number of passengers on bus i, and p_i be the event that a random person is on bus i. By Bayes' theorem

$$P(N_i|p_i) = \frac{P(p_i|N_i)P(N_i)}{P(p_i)}.$$

A randomly sampled passenger has a probability $P(p_i|N_i) = N_i/N$ of being on bus i if it has N_i passengers, where N is the total number of passengers: i.e. if there are more passengers on a bus the probability that a random person is on it is higher. Also $P(p_i) = 1/N_B$ where N_B is the total number of buses. Hence

$$P(N_i|p_i) = \frac{N_i N_B}{N} P(N_i).$$

So calculating the average

$$\sum_{N_i} N_i P(N_i|p_i) = \frac{N_B}{N} \sum_{N_i} N_i^2 P(N_i).$$

Approximate N_i as a continuous variable, with people arriving at a rate w, so $N_i = wt_i$. Now $P(t_i) = \nu e^{-\nu t_i}$, so

$$\langle N_i \rangle = \int N_i P(N_i) dN_i = \int N_i P(t_i) dt_i = \int w t_i \nu e^{-\nu t_i} dt_i = \frac{w}{\nu}.$$

$$\langle N_i \rangle_{\rm p_i} = \frac{N_B}{N} \int N_i^2 P(N_i) dN_i = \frac{1}{\langle N_i \rangle} \int N_i^2 P(t_i) dt_i = \frac{\nu}{w} \int t_i^2 w^2 \nu e^{-\nu t_i} dt_i = \frac{2w}{\nu}.$$

Hence the expected number of passengers given that you were a passenger is twice the average number of passengers in each bus. People usually see buses significantly fuller than the average bus (there's still a selection effect even if buses run on a more reliable schedule).
Follow up: why do cars often seem to travel faster in the other lane?