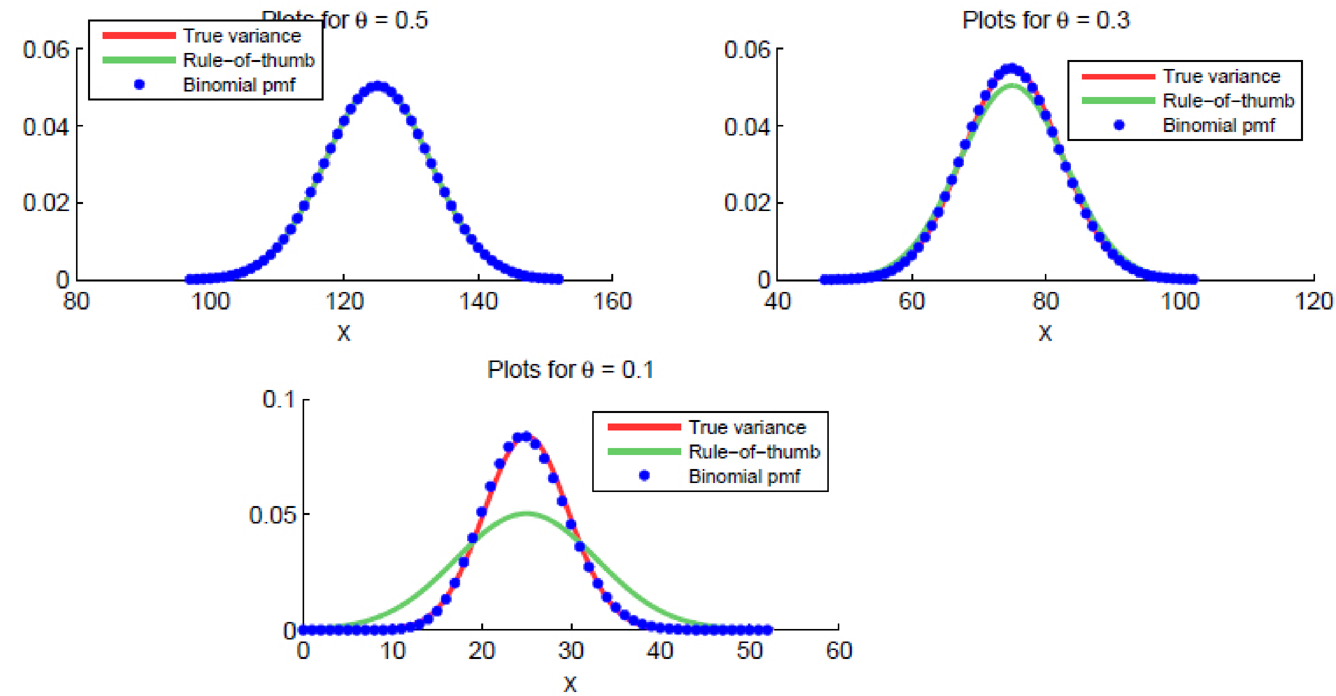
18.05 Problem Set 9, Spring 2014 Solutions

Problem 1. (10 pts.) **(a)** We have $x \sim \text{binomial}(n, \theta)$, so $E(X) = n\theta$ and $\text{Var}(X) = n\theta(1-\theta)$. The rule-of-thumb variance is just $\frac{n}{4}$. So the distributions being plotted are

$$\text{binomial}(250,\ \theta), \quad \text{N}(250\theta,\ 250\theta(1-\theta)), \quad \text{N}(250\theta,\ 250/4).$$

Note, the whole range is from 0 to 250, but we only plotted the parts where the graphs were not all 0.



We notice that for each θ the blue dots lie very close to the red curve. So the $\text{N}(n\theta, n\theta(1-\theta))$ distribution is quite close to the $\text{binomial}(n, \theta)$ distribution for each of the values of θ considered. In fact, this is true for all θ by the Central Limit Theorem. For $\theta = 0.5$ the rule-of-thumb gives the exact variance. For $\theta = 0.3$ the rule-of-thumb approximation is very good: it has smaller peak and slightly fatter tail. For $\theta = 0.1$ the rule-of-thumb approximation breaks down and is not very good.

In summary we can say two things about the rule-of-thumb approximation:
1. It is good for θ near 0.5 and breaks down for extreme values of θ. 2. Since the rule-of-thumb overestimates the variance (the rule-of-thumb graphs are shorter and wider) it gives us a confidence interval that is larger than is srictly necessary. That is a 95% rule-of-thumb interval actually has a greater than 95% confidence level.

(b) Using the rule-of-thumb approximation, we know that $\bar{x}$ is approximately $\text{N}(\theta, 1/4n)$. For an 80% confidence interval, we have $\alpha = 0.2$ so

$$z_{\alpha/2} = \text{ qnorm(0.9,0,1)} = 1.2815.$$

So the 80% confidence interval for θ is given by

$$\left[\bar{x} - \frac{z_{0.1}}{2\sqrt{n}}, \bar{x} + \frac{z_{0.1}}{2\sqrt{n}}\right] = [0.5195, 0.6005]$$

For the 95% confidence interval, we use the rule-of-thumb that $z_{0.025} \approx 2$. So the confidence interval is

$$\left[\bar{x} - \frac{1}{\sqrt{n}}, \bar{x} + \frac{1}{\sqrt{n}}\right] = [0.497, 0.623]$$

It's okay to have used the exact value of $z_{0.025}$. This gives a confidence interval:

$$\left[\bar{x} - \frac{1.96}{2\sqrt{n}}, \bar{x} + \frac{1.96}{2\sqrt{n}}\right] = [0.498, 0.622]$$

(c) With prior Beta(1, 1), if observe x and then the posterior is Beta$(x + 1, 250 + 1 - x)$. In our case $x = 140$. So, using R we get the 80% posterior probability interval:

$$\begin{aligned} \text{prob\_interval} &= [\texttt{qbeta(0.1, 141, 111), qbeta(0.9, 141, 111)}] \\ &= [0.51937,\ 0.5995] \end{aligned}$$

This is quite close to the 80% confidence interval. Though the two intervals have **very different** technical meanings, we see that they are consistent (and numerically close). Both give a type of estimate of θ.

Problem 2. (10 pts.) **(a)** We have $n = 20$ and $\alpha = 0.1$ so

$$t_{\alpha/2} = \texttt{qt(0.05,19) = 1.7291.}$$

Thus the 90% t-confidence interval is given by

$$\left[\overline{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}},\ \overline{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}\right] = [68.08993,\ 71.01007]$$

Given that the sample mean and variance are only reported to 2 decimal places the extra digits are a spurious precision. It is worth noting that to the given precision the 90% confidence interval is [68.08, 71.02]. (The problem did not ask you to do this.)

(b) We have

$$z_{\alpha/2} = \texttt{qnorm(0.05) = 1.6448.}$$

So the 90% z-confidence interval is given by

$$\left[\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},\ \overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right] = [68.15839,\ 70.94161]$$

As in part (a) taking the precision of the mean into account we get the interval [68.16, 70.94].

(c) We need n such that $2 \cdot z_{0.05} \cdot \sigma/\sqrt{n} = 1$. So $n = (2 \cdot z_{0.05} \cdot \sigma)^2 = 153.8$. Since you need a whole number of people the answer is $\boxed{n = 154}$.

(d) We need to find n so that $2 \cdot t_{0.05}/\sqrt{n} = 1$. Because the critical value $t_{0.05}$ depends on n the only way to find the right n is by systematically checking different values of n.

```
n = 157
t05 = qt(0.95,n-1) = 1.6547
width = (2*sqrt(s2)*t05/sqrt(n)) = 0.99736 (very close to 1).
```

(Our actual code used a 'for loop' to run through the values $n = 130$ to $n = 180$ and print the width to the screen for each n.)

We find $n = 157$ is the first value of n where the width 90% interval is less than 1. This is not guaranteed. In an actual experiment the value of s^2 won't necessarily equal 14.26. If it happens to be smaller then then the 90% t confidence interval will have width less than 1.

Problem 3. (10 pts.) **(a)** The sample mean is $\bar{x} = 356$. Since $z_{0.025} = 1.96$, $\sigma = 3$ and $n = 9$, the 95% confidence interval is

$$\left[\bar{x} - z_{0.025} \cdot \frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{0.025} \cdot \frac{\sigma}{\sqrt{n}}\right] = [354.04,\ 357.96]$$

(b)

We have $z_{0.01} =$ `qnorm(0.99)` $= 2.33$. So the 98% confidence interval is

$$\left[\bar{x} - z_{0.01} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{0.01} \cdot \frac{\sigma}{\sqrt{n}}\right] = [353.67, 358.33].$$

(c) The sample variance is

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \texttt{var([352 351 361 353 352 358 360 358 359])} = 15.5$$

Since $n = 9$ the number of degrees of freedom for the t-statistic is 8.

Redo (a): $t_{8,0.025} =$ `qt(0.975, 8)` $= 2.306$. So the 95% confidence interval is

$$\left[\bar{x} - t_{8,0.025} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{8,0.025} \cdot \frac{s}{\sqrt{n}}\right] \approx [352.97, 359.03].$$

Redo (b): $t_{8,0.01} =$ `qt(0.99, 8)` $= 2.896$. So the 98% confidence interval is

$$\left[\bar{x} - t_{8,0.01} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{8,0.01} \cdot \frac{s}{\sqrt{n}}\right] \approx [352.20, 359.80].$$

These intervals are larger than the corresponding intervals from parts (a) and (b). The new uncertainly regarding the value of σ means we need larger intervals to achieve the same level of confidence. This is reflected in the fact that the t distribution has fatter tails than the normal distribution).

Problem 4. (10 pts.) **(a)** This is similar to problem 3c. We assume the data is normally distributed with unknown mean μ and variance σ^2. We have the number of data points $n = 12$. Using Matlab we find

$$\begin{aligned}
\texttt{data} &= \texttt{[6.0, 6.4, 7.0, 5.8, 6.0, 5.8, 5.9, 6.7, 6.1, 6.5, 6.3, 5.8];} \\
\bar{x} &= \frac{\sum x_i}{n} = \texttt{mean(data)} = 6.1917 \\
s^2 &= \frac{\sum (x_i - \bar{x})^2}{n-1} = \texttt{var(data)} = 0.15356 \\
c_{0.025} &= \texttt{qchisq(0.975, 11)} = 21.920 \\
c_{0.975} &= \texttt{qchisq(0.025, 11)} = 3.8157
\end{aligned}$$

So the 95% confidence interval is

$$\frac{(n-1)\cdot s^2}{c_{0.025}}, \frac{(n-1)\cdot s^2}{c_{0.975}} = [0.077060, 0.442683].$$

s^2 is our point estimate for σ^2 and the confidence interval is our range estimate with 95% confidence.

(b) We have assumed that the plasma cholesterol levels are independent and normally distributed. This might not be a good assumption because cholesterol for men and women might follow different distributions. We'd have to do further exploration to understand this.

Problem 5. (10 pts.) **(a)** We have $n = 10$ and $s^2 = 4.2$ Assuming that the weights are normally distributed with mean $\mu = 52$ and variance σ^2, we know that $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_9$. We have

$$\begin{aligned} c_{0.025} &= \texttt{qchisq(0.975, 9)} = 19.023 \\ c_{0.975} &= \texttt{qchisq(0.025, 9)} = 2.7004 \end{aligned}$$

The 95% confidence interval for σ is given by

$$\left[\sqrt{\frac{s^2(n-1)}{c_{0.975}}}, \sqrt{\frac{s^2(n-1)}{c_{0.025}}}\right] = [1.4096, 3.7414]$$

(b) In order to use a χ^2 confidence interval we assumed that the weights of the packs of candy are independent and normally distributed with mean 52 and variance σ^2.