a) Revise the simulation shown in the lecture with the aim of constructing the empirical sampling distribution of beta_hat, based on 5000 trials.

```r
set.seed(123)     # Not necessary.
n = 10
ntrial = 5000
sigma_eps = 15
beta_hat = numeric(ntrial)
x = c(1:n)
for(trial in 1:ntrial){
y = 10 + 2*x + rnorm(n,0,sigma_eps)
lm.1 = lm(y~x)
alpha_hat[trial] = lm.1$coeff[1]
beta_hat[trial] = lm.1$coeff[2]
}
hist(beta_hat)
```

b) According to the lecture, the mean of that histogram is supposed to be approximately equal to the true slope. Is it? Show code.

```r
mean(beta_hat)      # 1.990397 i.e., close to 2.
```

c) According to the lecture, the standard deviation of that histogram is supposed to be approximately equal to sigma_eps/sqrt(Sxx). Is it? Show code.

```r
sd(beta_hat)                 # 1.623913 is close to
Sxx = sum((x-mean(x))^2)
sigma_eps/sqrt(Sxx)          # 1.651446
```

d) According to the lecture, the distribution of the beta_hat is supposed to be normal with certain parameters. Use qqnorm() and abline() to confirm that.

```r
qqnorm(beta_hat)
abline(2,sigma_eps/sqrt(Sxx), col=2) # y-int=true beta=2, slope = sigma_epsilon/sqrt(Sxx)
```

**hw_lect23_2**

In a problem dealing with flow rate (y) and pressure-drop (x) across filters, it is known that 
\[ y = -0.12 + 0.095 \, x \]  
Note: this is the true "fit" to the population. Suppose it is also known that 
\( \sigma_{\epsilon} = 0.025 \). Now, IF we were to make repeated observations of y when x=10,  
what's the prob. of a flow rate exceeding 0.835?

\[
P(y > 0.835 | x=10) = P\left( \frac{y - (\text{true mean of } y \text{ at } x=10)}{\sigma_{\epsilon}} > \frac{0.835 - (-0.12 + 0.095(10))}{0.025} \right)
\]

\[
= P(z > \frac{0.835 - 0.83}{0.025}) = 1 - 0.526793
\]

\[
= 0.47327
\]
Mist (airborne droplets or aerosols) is generated when metal-removing fluids are used in machining operations to cool and lubricate the tool and work-piece. Mist generation is a concern to OSHA, which has recently lowered substantially the workplace standard. The article "Variables Affecting Mist Generation from Metal Removal Fluids" (Lubrication Engr., 2002: 10Â17) gave the accompanying data on $x =$ fluid flow velocity for a 5% soluble oil (cm/sec) and $y =$ the extent of mist droplets having diameters smaller than some value:

$x$: 89 177 189 354 362 442 965
$y$: 0.40 0.60 0.48 0.66 0.61 0.69 0.99

a. Make a scatterplot of the data. By computer.

b. What is the point estimate the beta coefficient? (By computer) Interpret it.

c. What is $s_e$? (By computer.) Interpret it.

d. Estimate the true average change in mist associated with a 1 cm/sec increase in velocity, and do so in a way that conveys information about precision and reliability. Hint: This question is asking for a CI for beta. Compute it AND interpret it. By hand; i.e. you must use the basic formulas for the CI of beta: $\betahat + t \cdot s_e / \sqrt{S_{xx}}$, but you may use computer to compute the various terms in the formula.Use 95% confidence level.

e. Suppose the fluid velocity is 250 cm/sec. Compute an interval estimate of the corresponding mean $y$ value. Use 95% confidence level.

f. Suppose the fluid velocity for a specific fluid is 250 cm/sec. Predict the $y$ for that specific fluid in a way that that conveys information about precision and reliability. Use 95% prediction level.

```
x = c(89, 177, 189, 354, 362, 442, 965)
y = c(0.4, 0.6, 0.48, 0.66, 0.61, 0.69, 0.99)

# a)
plot(x,y)

# b) lm.1 = lm(y~x)
lm.1                     # alpha = 0.4041238  beta = 0.0006211
alpha_hat = 0.4041238
beta_hat = 0.0006211

# c)
s_emlm.1                   # s_e = 0.05405
s_e = 0.05405

# d) The
S_xx = (n-1)*var(x)

beta_hat - 2.571 * s_e/sqrt(S_xx)  # 2.571 (from Table IV) = qt(1-.05/2,n-2) in R
beta_hat + 2.571 * s_e/sqrt(S_xx)
c(0.0004262228, 0.0008159772)

# We are 95% confident that on the average a change of 1 unit in x will cause
# the mean of y to change by an amount between (0.00043, 0.00082)

# e)
y_hat = alpha_hat + beta_hat*250     # 0.5593988

s_y_hat = s_e * sqrt( (1/n) + ((250 - mean(x))^2)/S_xx )
y_hat - 2.571 * s_y_hat
y_hat + 2.571 * s_y_hat
c(0.5020402, 0.6167574)

# We are 95% confident that the true mean of y, at x = 250, is between (0.5020402, 0.6167574)

# f)
y_hat - 2.571 * sqrt(s_y_hat^2 + s_e^2)
y_hat + 2.571 * sqrt(s_y_hat^2 + s_e^2)
c(0.4090638, 0.7097338)

# About 95% of random PIs will cover the observed value of y, at x = 250.
```
a) Show that, at a given \( \alpha \), the probability that a randomly chosen \( y^* \) would fall within the observation P.I. for \( y^* \) is

\[
Pr\left( t_{obs} - t^* \sqrt{1 + \frac{1}{n} + \frac{(x-x)^2}{s_{xy}}} < t < t_{obs} + t^* \sqrt{1 + \frac{1}{n} + \frac{(x-x)^2}{s_{xy}}} \right) = \frac{\hat{y}_{obs}(x) - \hat{y}(x)}{s_{e}}
\]

(Hint: How do you standardize observation error?)

Obs P.I. for \( y^* \) : 

\[
\hat{y}(x) \pm t^* S_{pred.evr}
\]

\[
Pr\left( \frac{\hat{y}_{obs}(x) - t^* S_{pred.evr} < \hat{y} < \hat{y}_{obs}(x) + t^* S_{pred.evr}}{S_{e}} \right) = Pr\left( \frac{\hat{y}_{obs}(x) - \hat{y}(x) - t^* S_{pred.evr}}{S_{e}} < \frac{\hat{y}(x) - \hat{y}(x) + t^* S_{pred.evr}}{S_{e}} \right)
\]

\[
= Pr\left( \frac{t_{obs} - t^* S_{pred.evr}}{S_{e}} < t < + \right)
\]

\[
\frac{S_{e}}{\sqrt{S_{e}^2 + S_{0}^2}} = \frac{1}{S_{e}} \sqrt{S_{e}^2 \left[ \frac{1}{n} + \frac{(x-x)^2}{s_{xy}} \right] + S_{0}^2} = \sqrt{1 + \frac{1}{n} + \frac{(x-x)^2}{s_{xy}}}
\]

b) Show that, at a given \( \alpha \), the probability that a randomly chosen \( \hat{y}(x) \) would fall within the estimation P.I. for \( y^* \) is

\[
Pr\left( t_{obs} - t^* \sqrt{1 + \frac{1}{n} + \frac{(x-x)^2}{s_{xy}}} < t < t_{obs} + t^* \sqrt{1 + \frac{1}{n} + \frac{(x-x)^2}{s_{xy}}} \right) = \frac{\hat{y}_{obs}(x) - \hat{y}(x)}{S_{est.evr}}
\]

(Hint: How do you standardize estimation error?)

\[
Pr\left( \frac{\hat{y}_{obs}(x) - t^* S_{pred.evr} < \hat{y}(x) < \hat{y}_{obs}(x) + t^* S_{pred.evr}}{S_{est.evr}} \right)
\]

\[
= Pr\left( \frac{\hat{y}_{obs}(x) - \hat{y}(x) - t^* S_{pred.evr}}{S_{est.evr}} < \frac{\hat{y}(x) - \hat{y}(x) + t^* S_{pred.evr}}{S_{est.evr}} \right)
\]

\[
= Pr\left( \frac{t_{obs} - t^* S_{pred.evr}}{S_{est.evr}} < t < + \right)
\]

\[
\sqrt{\frac{S_{est.evr}^2 + S_{0}^2}{S_{est.evr}^2}} = \sqrt{1 + \frac{S_{e}^2}{S_{est.evr}^2} \left[ \frac{1}{n} + \frac{(x-x)^2}{s_{xy}} \right]} = \sqrt{1 + \frac{1}{n} + \frac{(x-x)^2}{s_{xy}}}
\]
Consider the defining formulas for C.I. and P.I.:

\[
\text{C.I. } \hat{\gamma}(x) \pm t^* \hat{S}_e \sqrt{\frac{1}{n} + \frac{(x-x)\hat{S}_{xx}}{S_x^2}} \quad \text{where } \hat{\gamma}(x) = \hat{\alpha} + \hat{\beta}x \\
\text{P.I. } \hat{\gamma}(x) \pm t^* \hat{S}_e \sqrt{\frac{1}{n} + \frac{(x-x)^2}{S_x^2}}
\]

a) As \( n \) becomes large (but not quite \( \infty \)) what does each of the following approach? For example \( \hat{\alpha} \to \alpha \).

\( \hat{\alpha} \to \alpha \) As \( n \) increases, \( \hat{\alpha} \) approaches the population param. \( \alpha \).

\( \hat{\beta} \to \beta \) As \( n \) increases, \( \hat{\beta} \) approaches the population param. \( \beta \).

\( \hat{\gamma}(x) \to \gamma(x) \) Pop. predicted value.

\( t^* \to z^* \) \( t \)-dist. approaches \( z \)-dist.

\( \hat{S}_e \to \sigma_e \) \( \hat{S}_e \) true std. dev. of errors.

\( \frac{\hat{x}}{\bar{x}} \) true mean of \( \alpha \).

\( S_{xx} = (n-1)S_x^2 \to \frac{(n-1)\sigma_x^2}{\sigma_x^2} \to (n-1) \) true var. of \( \hat{x} \).

\( \frac{\hat{x}}{\bar{x}} \to 1, \) as \( n \to \infty \)

b) As \( n \to \infty \), what does C.I. converge to?

\[
\hat{\gamma}(x) \pm t^* \hat{S}_e \sqrt{\frac{1}{n} + \frac{(x-x)\hat{S}_{xx}}{S_x^2}} \to \gamma(x) \pm z^* \hat{S}_e \sqrt{\frac{1}{n} + \frac{(x-x)^2}{\sigma_x^2}} = \gamma(x)
\]

c) As \( n \to \infty \), what does P.I. converge to?

\[
\hat{\gamma}(x) \pm t^* \hat{S}_e \sqrt{\frac{1}{n} + \frac{(x-x)^2}{S_x^2}} \to \gamma(x) \pm z^* \hat{S}_e \sqrt{\frac{1}{n} + \frac{(x-x)^2}{\sigma_x^2}}
\]

= \( \gamma(x) \pm z^* \hat{S}_e \).

Note. This makes sense. As \( n \to \infty \), you are looking at the whole pop. and so the C.I. collapses to just \( \gamma(x) \). However, the P.I. becomes \( \gamma(x) \pm z^* \hat{S}_e \) which is exactly the interval that covers 95% (for example) of the \( y \)'s, just like a P.I. is supposed to do.