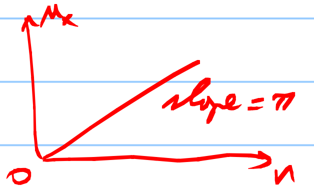
hw_lect9_1

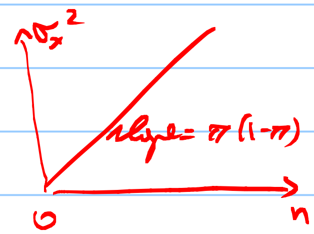
a) Consider the binomial dist. with params n,pi. Draw four figures that show qualitatively how its mean (mu_x) and variance (sigma_x^2) vary with n and pi.

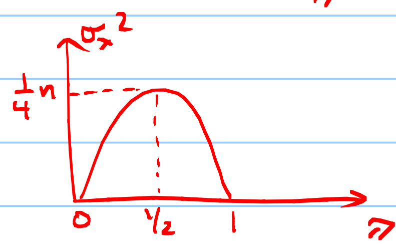
Suppose we toss n=l00 unfair coins, with an unknown pi.
b) What is the expected number of heads out of n? (The answer depends on pi),
c) What is the typical deviation in the number of heads out of n? (The answer depends on pi).
d) what is the largest typical deviation of the number of heads out of n? (The answer is a number!)
Hint: Consult your graph of variance vs. pi, in part a).

a) $\mu_x = n\pi \Rightarrow$



$\sigma_x^2 = n\pi(1-\pi) \Rightarrow$





b) $100\,\pi$

c) $\sqrt{100(\pi)(1-\pi)}$

d) The largest σ_x^2 (i.e. peak of the curve in σ_x^2 vs. π) occurs at $\pi = \frac{1}{2}$, and it is $100\frac{1}{2}(1-\frac{1}{2}) = 25$. So, the max dev. is 5.

This is an important property of binomial dist. Remember it!

hw-lect9-2

For the exponential distribution with parameter lambda, find the variance.
Hint: You may use the integral:

$$\int_0^\infty (y-1)^2 e^{-y}\, dy = 1$$

For exponential distr: $f(x) = \lambda e^{-\lambda x}$. Then,

$$\mu_x = E[x] = \int_{-\infty}^{\infty} x\, f(x)\, dx = \int_0^\infty \lambda x e^{-\lambda x}\, dx$$

$y = \lambda x$
$dy = \lambda dx$

$$= \frac{1}{\lambda}\int_0^\infty y e^{-y}\, dy = \boxed{\frac{1}{\lambda}}.$$

past hw. For your convenience it's here too.

$$\sigma_x^2 = V[x] = \int_{-\infty}^{\infty} (x-\mu_x)^2 f(x)\, dx = \int_0^\infty \left(x-\frac{1}{\lambda}\right)^2 \lambda e^{-\lambda x}\, dx$$

$$= \frac{1}{\lambda^2}\int_0^\infty (\lambda x - 1)^2 e^{-\lambda x}\, d(\lambda x)$$

$y = \lambda x$

$$= \frac{1}{\lambda^2}\underbrace{\int_0^\infty (y-1)^2 e^{-y}\, dy}_{=1} = \frac{1}{\lambda^2} \Rightarrow \boxed{\sigma_x = \frac{1}{\lambda}}$$

hw_lect9_4

Find the prob. that x is within 1 std-dev. of the mean, for

a) Binomial(n=20, pi=1/4) Binomial $(n=20, \pi=\frac{1}{4})$

b) Poisson(lambda=5) Poisson $(\lambda=5)$

c) Normal(mu=5, sigma=1) Normal $(\mu=5, \sigma=1)$

a) binomial $(n=20, \pi=\frac{1}{4})$

$\mu_x = n\pi = 20(\frac{1}{4}) = 5$, $\sigma_x = \sqrt{n\pi(1-\pi)} = \sqrt{20(\frac{1}{4})(\frac{3}{4})} = 1.9$

$\mu_x - \sigma_x < x < \mu_x + \sigma_x \Rightarrow 5-1.9 < x < 5+1.9 \Rightarrow 3.1 < x < 6.9$

Integer x values in this range are $x = 4, 5, 6$,

$\therefore$ area $= p(x=4) + p(x=5) + p(x=6) = 0.19 + .202 + .169 = \underline{0.561}$

Table II.

b) poisson $(\lambda=5)$

$\mu_x = \lambda = 5$, $\sigma_x = \sqrt{\lambda} = \sqrt{5} = 2.24$

$\mu_x - \sigma_x < x < \mu_x + \sigma_x \Rightarrow 5-2.24 < x < 5+2.24 \Rightarrow 2.76 < x < 7.24$

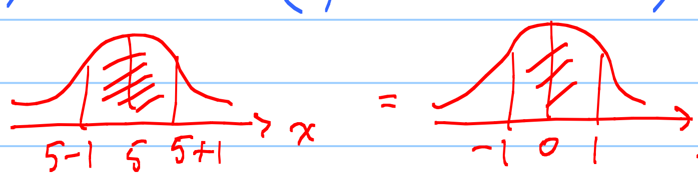
Integer x values in this range are $x = 3, 4, 5, 6, 7$.

$\therefore$ area $= p(x=3) + \cdots + p(x=7) = 0.14 + 0.175 + 0.175 + 0.146 + 0.104$

Table III.

$= \underline{0.742}$

c) Normal $(\mu=5, \sigma=1)$



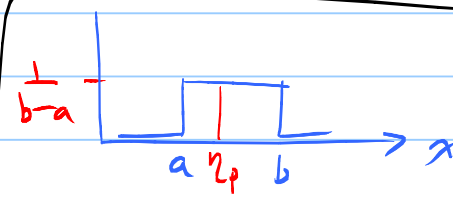
$z = \frac{x-5}{1}$

Table I

$= .8413 - .1587 = \underline{.6826}$

hw-lect9-4

It can be shown that the p^{th} percentile of Unif (a,b) is given by $\eta_p(a,b) = a + (b-a)\frac{p}{100}$



$$\frac{(\eta_p - a)}{b-a} = \frac{p}{100}$$

$$\eta_p = a + (b-a)\frac{p}{100}$$

a) What's the p^{th} percentile of Unif$(0,1)$, ie. $\eta_p(0,1)$?

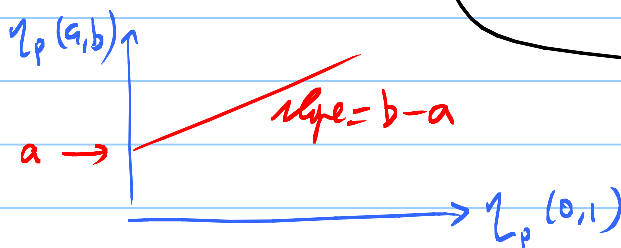
$$\eta_p(0,1) = 0 + \frac{p}{100}$$

b) Write $\eta_p(a,b)$ in terms of $\eta_p(0,1)$.

$$\eta_p(a,b) = a + (b-a)\,\eta_p(0,1)$$

c) What will the plot of $\eta_p(a,b)$ vs. $\eta_p(0,1)$ look like? What are the slope & y-intercept?

A straight line with y-intercept a, and slope $(b-a)$.



Moral: If you make a qq plot of your data, but with $\eta_p(0,1)$ on the x-axis, then a straight line would imply that your data come from a unif(a,b) with a, b estimated <u>from</u> y-intercept & slope $(b-a)$.

hw_lect9_5

Do a qq-plot of each of the 2 cont. vars. in the data you collected. By R. Describe/Interpret the result.
Note: If you find out that there is not much you can say about the qq-plot, it may be that your data is not appropriate. This is another chance to correct the error, because later you will be doing more hw problems using your data. So, see me, if you are not sure.

The results will vary across students, but if you see a "straight" qq plot, Then That's evidence That your data came from a Normal distr. with μ equal to The y-intercept, and σ = slope of The qq plot. If your qq plot looks curved, Then try doing The qq plot of The log or $\sqrt{}$ of The data. If that makes The qq plot "straight" Then use The log or $\sqrt{}$ of The data for all future work.

hw-lect10-1 Same comments as

hw_lect10_2

Suppose n cases of data on x and y fall exactly on the line y = mx + b. Compute the value of r.
Hint: In any of the formulas for r, eliminate all y's in favor of x's.

$\bar{y} = m\bar{x}$

$$r = \frac{1}{n-1}\sum_{i}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) \qquad s_y^2 = m^2 s_x \Rightarrow s_y = |m| s_x$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{m x_i \cancel{+b} - m\bar{x} \cancel{-b}}{|m| s_x}\right)$$

$$= \frac{m}{|m|}\frac{1}{s_x^2}\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{m}{|m|}\frac{s_x^2}{s_x^2} = \pm 1 \qquad (\text{if } m \neq 0)$$

note that The magnitude of m, The slope, cancels out (if it's not 0, ∞). So, The magnitude of The slope is irrelevant to r.

hw-lect10-3 I gave you a formula that defines r. The book gives two others on p. 110.

a) Start from the formula I "derived" in class, and show that it is equal to
$$r = \frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_j (x_j-\bar{x})^2}\sqrt{\sum_k (y_k-\bar{y})^2}} \qquad \textcircled{\text{I}}$$

$$r = \frac{1}{n-1}\sum_i \left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

s_x, s_y have no $\underline{i}$ index.

$$r = \frac{1}{s_x s_y}\frac{1}{n-1}\sum_i (x_i-\bar{x})(y_i-\bar{y})$$

$$= \frac{1}{\sqrt{\cancel{\frac{1}{n-1}}\sum_j (x_j-\bar{x})^2}} \cdot \frac{1}{\sqrt{\cancel{\frac{1}{n-1}}\sum_k (y_k-\bar{y})^2}} \cdot \cancel{\frac{1}{n-1}}\sum_i (x_i-\bar{x})(y_i-\bar{y})$$

Defn. of s_x, s_y

$$= \frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_j (x_j-\bar{x})^2}\sqrt{\sum_k (y_k-\bar{y})^2}}$$

Use different indices so that you don't cancel $(x_i-\bar{x})$ in the numerator with that in the denominator.

b) In the lecture where we "derived" the formula for the sample variance, we introduced $S_{xx} \equiv \sum_i (x_i-\bar{x})^2$, i.e. the numerator of sample variance. If we similarly define S_{yy}, and $\underline{S_{xy} = \sum_i (x_i-\bar{x})(y_i-\bar{y})}$, then it follows from (I) that $r = S_{xy}/\sqrt{S_{xx}}\sqrt{S_{yy}}$ (which is the other formula for $\underline{r}$ on p. 110).

To finish all the comparison between my formula for $\underline{r}$ and those in the book, show that S_{xy} as defined here is equal to $\sum_i x_i y_i - n\bar{x}\bar{y}$ (or $n(\overline{xy} - \bar{x}\bar{y})$) as written in the book.

Hint: Recall $\sum_i (y_i-\bar{y}) = 0$ [or $\sum_i (x_i-\bar{x}) = 0$].

$$S_{xy} = \sum_i (x_i-\bar{x})(y_i-\bar{y}) = \sum_i (x_i-\bar{x})y_i - \sum_i (x_i-\bar{x})\bar{y}$$

$$= \underbrace{\sum x_i y_i}_{n\overline{xy}} - \bar{x}\underbrace{\sum y_i}_{n\bar{y}} - \bar{y}\underbrace{\sum (x_i-\bar{x})}_{=0} = n(\overline{xy} - \bar{x}\bar{y})$$

As for S_{xx} and S_{yy}, defined above, we have already shown that $S_{xx} \equiv \sum (x_i-\bar{x})^2 = n(\overline{x^2} - \bar{x}^2)$; see the sample var. lecture.

hw_lect11_1

I encourage you to come-up with other examples with "Easy" and "Hard" variables, because you're likely to come across something practically useful. But, even if you don't want to do that, do develop a regression model for predicting one of the continuous variables in your collected data from the other continuous variable. By R. Include your code, and report and interpret the slope parameter.

The answers will vary across students (hopefully); so, here I'm just going to repeat the example I discussed in class:

Consider the problem of measuring Intracranial Pressure (ICP), which is the pressure inside the brain - a quantity that's very important for doctor when a patient is brought into an emergency room. Ordinarily, the way it's measured is by drilling a hole in the head, inserting a long probe into the brain, and then making the measurements. That's clearly a "hard" thing to do. Now, suppose blood pressure measured easily on the arm is correlated with ICP. Or suppose we have an Ultrasound Doppler radar thatwe can point at the brain and measure the Flow of blood in to the brain. This is clearly "easy" to do because it requires no surgery. After one collects data on Flow and ICP for a bunch of patients, and develops the regression model, then one doesn't have to drill heads any more! One can simply use the easy thing (Flow) to predict the hard thing (ICP).

hw_lect11_2: Show that $\frac{\partial}{\partial\alpha} MSE\Big|_{\hat\alpha,\hat\beta} = 0$ implies $\bar{y} - \hat\alpha - \hat\beta\bar{x} = 0$

$$MSE(\alpha,\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$$

$$\frac{\partial}{\partial\alpha} MSE(\alpha,\beta) = \frac{1}{n}\sum_{i=1}^{n}\underbrace{\frac{\partial}{\partial\alpha}(y_i-\alpha-\beta x_i)^2}_{2(y_i-\alpha-\beta x_i)^1\cdot\underbrace{\frac{\partial}{\partial\alpha}(-\alpha)}_{-1}}$$

chain rule.

$$= \frac{2}{n}\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)$$

$$= \frac{2}{n}\Big(\underbrace{\sum_i^n y_i}_{n\bar{y}} - \alpha\underbrace{\sum_i^n 1}_{n} - \beta\underbrace{\sum_i^n x_i}_{n\bar{x}}\Big) = 2(\bar{y} - \alpha - \beta\bar{x})$$

$$\frac{\partial}{\partial\alpha} MSE\Big|_{\hat\alpha,\hat\beta} = 0 \implies \underline{\bar{y} - \hat\alpha - \hat\beta\bar{x} = 0}$$

---

hw_lect11-3 Prove that the Ordinary Least Square (OLS) fit, (i.e. the one described in this lecture) goes through the point $(\bar{x},\bar{y})$.
Hint: All you need is the Normal eqn. for $\hat\alpha$.

At $x=\bar{x}$, we have $\hat{y} = \hat\alpha + \hat\beta\bar{x} = \bar{y} - \cancel{\hat\beta\bar{x}} + \cancel{\hat\beta\bar{x}} = \bar{y}$

I.e. The OLS fit always goes thru the point $(\bar{x},\bar{y})$.

hw-lect11-4 : Show that $\hat{\beta}$ as defined by $\frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}$ or $\frac{S_{xy}}{S_{xx}}$ can be written as $\hat{\beta} = r\frac{s_y}{s_x}$ where s_x = sample std. dev. of x.
s_y = " " " " y.

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{S_{xy}}{S_{xx}}\frac{S_{yy}}{S_{yy}} = \frac{S_{xy}}{S_{xx}S_{yy}} S_{yy}$$

$$S_{xx}S_{yy} = \sqrt{S_{xx}S_{yy}}\sqrt{S_{xx}S_{yy}}$$

$$= \underbrace{\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}}_{r} \cdot \frac{S_{yy}}{\sqrt{S_{xx}S_{yy}}}$$

$$= r \cdot \sqrt{\frac{S_{yy}}{S_{xx}}} = r \cdot \sqrt{\frac{(n-1)s_y^2}{(n-1)s_x^2}} = \boxed{r \cdot \frac{s_y}{s_x}}$$

$\pm$

hw_lect11_5
Values of modulus of elasticity (MoE, the ratio of stress, i.e., force per unit area, to strain, i.e., deformation per unit length, in GPa) and flexural strength (a measure of the ability to resist failure in bending in MPa) were determined for a sample of concrete beams of a certain type, resulting in the following data (read from a graph in the article "Effects of Aggregates and Microfillers on the Flexural Propertie of Concrete," Magazine of Concrete Research, 1997 8198):
MoE:
29.8 33.2 33.7 35.3 35.5 36.1 36.2 36.3 37.5 37.7 38.7 38.8 39.6 41.0 42.8 42.8 43.5 45.6 46.0 46.9 48.0 49.3 51.7 62.6 69.8 79.5 80.0
Strength:
5.9 7.2 7.3 6.3 8.1 6.8 7.0 7.6 6.8 6.5 7.0 6.3 7.9 9.0 8.2 8.7 7.8 9.7 7.4 7.7 9.7 7.8 7.7 11.6 11.3 11.8 10.7

a) Plot a scatterplot of Strength vs. MOE. By R.
b) Make a boxplot of MOE, and of Strength. By R.
c) Make a qqplot of MOE, and of Strength. By R.
d) Compute the correlation coefficient between MOE and Strength. By hand. You may use the computer to compute sample means of necessary quantities,but you must use one of the formulas for r.
e) Compare it with the correlation you get from cor() in R.
f) Compute the equation of the OLS fit (i.e., the intercept and slope). By hand.You may use the computer to compute sample means of necessary quantities,but you must use the formulas for OLS intercept and slope).
g) Interpret the slope.
h) Predict Strength when MoE is 39.0 . By hand.
i) Compute the sum squared error (SSE, or SSResid). By hand, but you may use R to compute sample means of necessary quantities.

```
MOE =
c(29.8,33.2,33.7,35.3,35.5,36.1,36.2,36.3,37.5,37.7,38.7,38.8,39.6,41.0,42.8,42.8,43.5,45.6,46.0,46.9,48.0,49.3,51.7,62.6,69.8,79.5,80.
0)

Strength = c(5.9,7.2,7.3,6.3,8.1,6.8,7.0,7.6,6.8,6.5,7.0,6.3,7.9,9.0,8.2,8.7,7.8,9.7,7.4,7.7,9.7,7.8,7.7,11.6,11.3,11.8,10.7)

# a)
plot(MOE, Strength)          # There is a decent linear association.

# b)
boxplot(MOE, Strength)       # No comment.

# c)
qqnorm(MOE)                  # Neither distribution is truly Normal, and so,
qqnorm(Strength)             # technically we should "fix it" before proceeding.

# d)
zx = (MOE - mean(MOE))/sd(MOE)
zy = (Strength - mean(Strength))/sd(Strength)
sum(zx*zy)/(length(MOE)-1)   # 0.859, i.e., decent correlation

# e)
cor(MOE,Strength)            # same as "by hand" above.

# f)
numerator = mean(MOE*Strength) - mean(MOE)*mean(Strength)
denominator =  mean(MOE^2) - (mean(MOE))^2
beta = numerator/denominator
beta                         # 0.1075
alpha = mean(Strength) - beta*mean(MOE)
alpha                        # 3.295

# g)
# For every unit change in MOE, on the average Strength changes by 0.1075 .
# FYI: When MOE = 0 , Strength is expected to be about 3.295. But note that MOE = 0 is really an extrapolation,
# because the range of MOE is 29 to 80.
range(MOE)

# h)
alpha + beta*39.0            # 7.48

# i)
yhat = alpha + beta*MOE
sum((Strength - yhat)^2)     # 18.735,  i.e., the smallest possible SSE.
```