Test 3 for MTH 113 will be November 24th. As agreed upon in class, this is a week later than scheduled. Test 4 is still scheduled for the 8th of december – only 4 class periods away. We are currently covering material that will appear on the fourth test.
As for test 3, the material on the exam will come from lessons 28,29,30,31,34,35,36 and 37. That is, pages 403-431 and 581-641.
Here are the main skills that will be tested on the exam:

The normal distribution We learned that there is a distribution of random number called the **normal distribution**. We learned that there are two *parameters* that dictate the shape of the distribution: the mean, μ, and the variance, σ^2. We learned that probabilities for the normal distribution are given by **areas** and we can calculate these areas using a table. In order to do so, we need to be able to convert a problem involving a normal with given μ and σ to one where $\mu = 1$ and $\sigma = 1$. This is from z-scores. the important formulas for this are

$$z = \frac{x - \mu}{\sigma}, \quad x = \mu + z\sigma.$$

That is, the z score is the number of standard deviations x is from the mean.

Once we have a z score, probabilities can be calculated for the corresponding area of the standard normal. All probabilities in the book are given in terms of an area to the left of z and right of 0. This means we usually have extra work to do to answer our problem.

Problems involving normal distributions There are two types of problems that involve normal distribution. First, many distributions in life are well described by the normal distribution. Examples are SAT scores for college students, the heights of individuals or the gestation times of newborns.

Other problems that involve the normal distribution come from the fact that the normal distributions can be used to give approximate answers to binomial problems. That is, if X is a binomial random variable with parameters n and p then we have

$$P(X \leq b) \approx P(Z < \frac{b - np}{\sqrt{np(1-p)}}).$$

That is, X is approximately normal with mean $\mu = np$ and variance $\sigma^2 = np(1-0)$.

The above equation does not include the **continuity correction**. This adds and subtracts a value of 1/2 as appropriate. This gives the formula

$$P(a \leq X \leq b) \approx P\left(\frac{a - \frac{1}{2} - np}{\sqrt{np(1-p)}} \leq Z \leq \frac{b + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right).$$

On the exam, you can answer with either **unless** I explicitly ask for one or the other.

To use the above, you need to do the following

- Observe that you are being asked about a binomial random variable
- Figure out the n and p for the binomial random variable
- Figure out the $\mu = np$ and $\sigma = \sqrt{np(1-p)}$ for the normal approximation
- Write your problem in terms of $P(a \le X)$, $P(X \le b)$ or $P(a \le X \le b)$ and then apply the normal approximation.
- Finally, you need to find the area for the normal.

Correlation We talked about relationships between two variables X and Y. For this we learned how to make a scatterplot. And how to find the **correlation.** This involved the formula

$$\text{correlation} = r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}.$$

We learned that we can guess a value for r from a scatterplot. A shotgun picture has a r value near 0, If the points are along a line, then r is close to 1 or -1. This is what correlated means. there is relationship between an x value and the corresponding y value.

Linear regression If the data appear to be along a straight line, then we might try to summarize the relationship with a regression line. The regression line is found from the data and is of the form

$$\begin{aligned} \hat{y} &= b_0 + b_1 x, \\ b_1 &= r\frac{s_y}{s_x} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}, \\ b_0 &= \bar{y} - b_1 \bar{x}. \end{aligned}$$

These come by applying the **method of least squares** to the data.

We can use the regression line to **predict** a value for y if we have a new value of x. That is $\hat{y} = b_0 + b_1 x$.

Statistical model for regression The statistical model for regression is that

$$y_i = \beta_0 + \beta_1 x_i + error_i.$$

The errors are assumed to be normal with mean 0 and variance σ^2. What this means is that for a given value of x, say x_i, the corresponding value of y_i is found by first multiplying x_i by β_1 and adding β_0, then we add a random value to this.

When we see the data, we try to find the line, by estimating β_0 and β_1 from the data with b_0 and b_1.

The value of σ^2 control how close to a line the data will be. We can estimate σ^2 with

$$s_e^2 = \frac{1}{n-2}\sum(y_i - \hat{y}_i)^2.$$

The latter is the sum of the squared residual errors. These were made as small as possible to find b_0 and b_1.

R^2 Finally, we noted that for the regression we have

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2.$$

This led to the following definition

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}.$$

This is close to 1 if $\sum (y_i - \hat{y}_i)^2$ is small. That is when the actual y values are close to the predicted ones. It is close to 0 when this isn't true.

One can show that $R^2 = r^2$ where the little r is the correlation coefficient we found previously.

1 Problems

1. Let X be normal with mean 60 and standard deviation 10. Find these **without** using a table
   (a) $P(50 \leq X \leq 70)$
   (b) $P(X < 80)$
   (c) $P(X > 70)$.
2. Let X be normal with mean 60 and standard deviation 10. FInd these using a table
   (a) $P(55 < X)$
   (b) $P(X < 45)$
3. If X is normal with mean 100 and variance 25, find the 25th percentile of X. That is find x so that $P(X < x) = 0.25$.
4. Assume the adult black bear has a normally distributed weight with mean 400 and standard deviation 25. Find the probability that a randomly chosen adult black bear is 375 or more pounds.
5. Toss a coin 400 times. Use the normal approximation to find the probability that your get 175 or fewer heads.
6. You have store front which has 10,000 people drive by everyday. If the probability that a random person driving by will stop by is 1 in 1000, use the normal approximation to find the probability of 12 or more customers in a day.

For the following, let the data be

```
x| 1 1 2 2 3 3
y| 1 2 1 2 3 3
```

1. Make a scatterplot of x and y.
2. Draw, by hand, your guess of the regression line. Find the slope and intercept by identifying two points on your line.
3. Find the correlation between x and y
4. Find the slope and intercept of the least-squares regression line.
5. Find the residual error for the point $(2, 1)$.
6. Find the predicted value for $x = 1.5$, $x = 2.5$

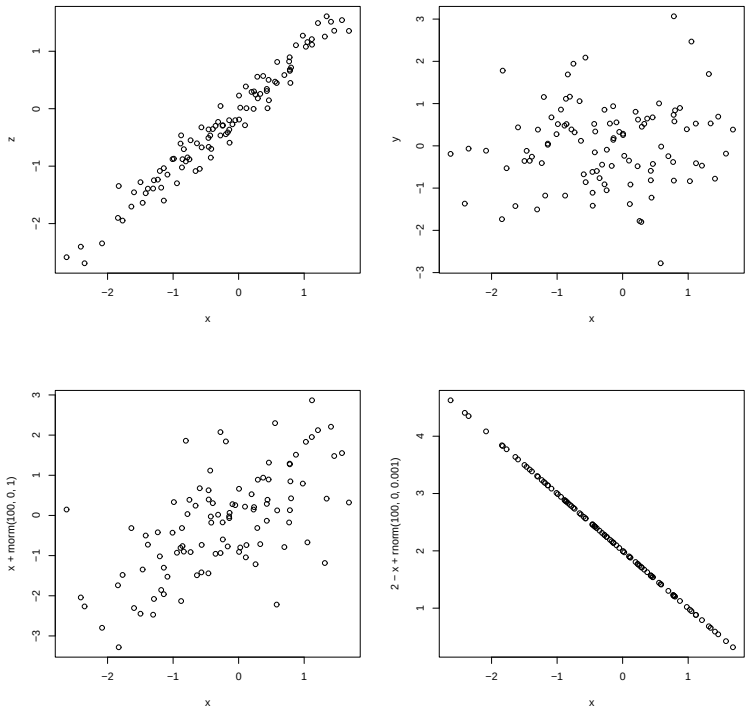

Figure 1: various scatterplots

For the figure answer

1. Which scatterplot has the largest R^2?
2. Which scatterplot has the smallets R^2?
3. which scatterplot has a negative value of r?
4. Draw by hand the 4 regression lines.