

22S:30/105  
Statistical Methods and  
Computing

Wrap-Up of Normal Distributions  
Intro to Regression

Lecture 5  
Jan. 30, 2013

Kate Cowles  
374 SH, 335-0727  
kate-cowles@uiowa.edu

z-scores tell how many standard deviations the original observation is away from the mean of the distribution, and in which direction.

- If the z-score is positive, the original observation was larger than the mean  $\mu$ .
- If the z-score is negative, the original observation was smaller than  $\mu$ .

Standardizing values from other normal distributions

All normal distributions would be the same if we measured in units of size  $\sigma$  around the mean  $\mu$  as center!

If  $x$  is an observation from a distribution that has mean  $\mu$  and standard deviation  $\sigma$ , the *standardized value* of  $x$  is

$$z = \frac{x - \mu}{\sigma}$$

Standardized values are often called *z-scores*.

Example of z-scores

Recall that the distribution of systolic blood pressure of men aged 18-74 is approximately normal with  $\mu = 129$  mm Hg and  $\sigma = 20$  mm Hg. The standardized height is

$$z = \frac{sbp - 129}{20}$$

If a man has  $sbp = 157$  mm Hg, his standardized sbp is

$$z = \frac{157 - 129}{20} = 1.4$$

If a man has  $sbp = 93$  mm Hg, his standardized sbp is

$$z = \frac{93 - 129}{20} = -1.8$$

## Using the standard normal distribution to compute proportions for other normal distributions

Let's use the symbol  $X$  for a variable representing the systolic blood pressure of men. What proportion of men have  $\text{sbp} < 100$ ?

If a man has  $\text{sbp} = 100$ , his standardized  $\text{sbp}$  is

$$z = \frac{100 - 129}{20} = -1.45$$

According to the normal table, the proportion of values of a standard normal variable that are less than or equal to  $-1.45$  is  $0.0735$ .

This proportion is the same as the proportion of  $X$  values that will be less than  $100$ .

## Example

For women in the US between 18 and 74 years of age, diastolic blood pressure follows a normal distribution with mean is  $\mu = 77$  mm Hg and standard deviation  $\sigma = 11.6$  mm Hg.

We want to know the proportion of US women in this age group who have  $\text{dbp}$  between  $60$  and  $100$ .

## General procedure for finding normal proportions

1. State the problem in terms of the observed variable  $X$ .
2. Standardize the value of interest  $x$  to restate the problem in terms of a standard normal variable  $Z$ . You may then wish to draw a picture to show the area under the standard normal curve.
3. Find the required area under the standard normal curve, using Table A and remembering
  - The total area under the curve is  $1.0$ .
  - The normal distribution is symmetric.

1. Call the variable representing a woman's  $\text{dbp}$   $X$ , and call the specific value for an individual woman  $x$ .  $X$  has a normal distribution with  $\mu = 77$  and  $\sigma = 11.6$ . We want to compute the proportion of women such that

$$60 \leq X \leq 100$$

2. Standardize  $x$  to produce  $z$ , a draw from a standard normal distribution.

$$60 \leq X \leq 100$$

$$\frac{60 - 77}{11.6} \leq \frac{X - 77}{11.6} \leq \frac{100 - 77}{11.6}$$

$$-1.47 \leq Z \leq 1.98$$

3. Use Table A to find
- the proportion of  $Z$  values  $\leq -1.47$ , which = .0708
  - and the proportion of  $Z$  values  $\leq 1.98$ , which = .9761.
4. So the percent of women with diastolic blood pressure between 60 and 100 is about  $97.61\% - 7.08\% = 90.5\%$ .

**General formula for *unstandardizing* a z-score:**

$$x = \mu + z\sigma$$

### Normal calculations going the other direction

What is the value of dbp such that 10% of women have values greater than or equal to it?

1. Use Table A to find the z-score such that 10% of a standard normal population would have values greater than or equal to it. This is the same value such that 90% of values are less than or equal to it, namely 1.28.
2. Convert  $z = 1.28$  into  $x$ .

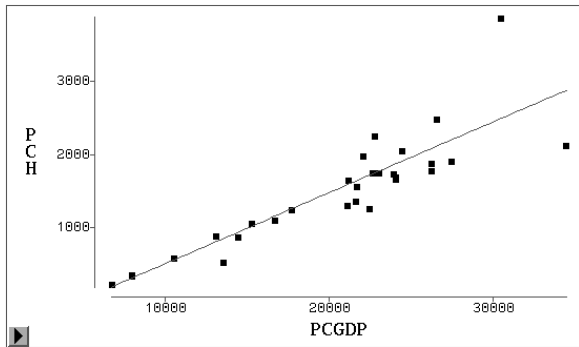
$$\begin{aligned} \frac{x - \mu}{\sigma} &= z \\ \frac{x - 77}{11.6} &= 1.28 \\ x &= 77 + (11.6)(1.28) \\ x &= 91.85 \end{aligned}$$

### Scatterplots

- represent the relationship between two different continuous variables measured on the same subjects
- each point represents the values for one subject for the two variables

Example: data reported by the Organization for Economic Development and Cooperation on its 29 member nations in 1998

- Per capita gross domestic product (a measure of wealth of the country) is on x-axis (horizontal)
- Per capita health care expenditures is on y-axis (vertical)



## Positive and negative association

- Two variables are *positively associated* when above-average values of one tend to occur in individuals with above-average values of the other, and below-average values of both also tend to occur together.
- Two variables are *negatively associated* when above-average values of one tend to occur in individuals with below-average values of the other, and vice-versa.

## We can describe the overall pattern of a scatterplot by

- form or shape
- direction
- strength

## Linear relationships

- The form of a relationship shown by a scatterplot is linear if the points lie in a straight-line pattern.
- The linear relationship is strong if the points lie close to a line, with little scatter.

### Example: per capita health care expenditures and gross domestic product

- “individuals” studied are countries
- form of relationship is roughly linear
- direction of relationship is positive
- strength: determined by how closely the points follow a clear pattern
  - quite strong

### Computing the sample correlation coefficient

- We have measured two different variables  $X$  and  $Y$  on the subjects in a study.
- There are  $n$  subjects.
- Let  $\bar{x}$  and  $\bar{y}$  be the sample means of the two variables.
- Denote the sample standard deviation of the  $x$  variable as  $s_x$  and the sample standard deviation of the  $y$  variable as  $s_y$ .
- Then the sample correlation coefficient is computed as

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

### Correlation

- a numeric measure of the direction and strength of the linear relationship between two continuous variables measured on the same subjects
- terminology and notation
  - sample correlation coefficient  $r$

- Note that the first step in computing  $r$  is to *standardize* the measurements.
- Example: suppose  $X$  is heart rate in beats per minute and  $Y$  is body temperature in degrees Fahrenheit, and we have both heart rate and temperature measurements on  $n = 10$  people.
  - The quantity

$$\frac{x_i - \bar{x}}{s_x}$$

is the standardized heart rate for person  $i$

\* how many standard deviations above or below the mean heart rate person  $i$ 's heart rate is

- Standardized values are no longer in their original units (e.g., the standardized heart rates are not in beats per minute)

- The sample correlation coefficient  $r$  is an average of the products of the standardized heart rates and temperatures for the 10 people.

- values near +1 indicate a very strong positive relationship (all points lie almost exactly on a straight line)
- values near -1 indicate a very strong negative relationship (all points lie almost exactly on a straight line)
- Correlation measures only the strength of *linear* relationships.  $r$  may be close to 0 even if the relationship between two variables is strong, if that relationship is curved.
- The sample correlation coefficient is very sensitive to outliers.
- A high correlation between two variables does not by itself imply a causal relationship.

## Facts about correlation

- Correlation requires that both variables be quantitative, so that we can do arithmetic computations with them.
- $r$  has no units, and, because it uses standardized values, it does not change when we change the units of measurements of  $x$ ,  $y$ , or both.
  - For the same 10 people,  $r$  would not change whether we measured the heights and weights in inches and pounds or in centimeters and kilograms.
- $r > 0$  indicates a positive association between the two variable;  $r < 0$  indicates a negative association
- $r$  is always between -1 and +1
  - values of  $r$  near 0 mean a very weak linear relationship

## Correlation and regression

- Correlation enables us to *assess the strength* of a linear relationship between two variables, but it does not enable us to *predict* the value of one variable for a subject for whom we know the value of the other variable.
- Prediction often is an important goal of statistical analysis.
- Example: we may wish to predict an infant's birthweight based on a laboratory measurement taken on the mother during pregnancy

## Response variables and explanatory variables

- **response variable**

- what we want to explain or predict
- also called “dependent” or “outcome” variable

- **explanatory variable**

- a variable that explains or influences differences in a response variable
- also called “predictor” variables, “covariates,” or “independent” variables

- When making a scatterplot of such data:

- response variable goes on y-axis (vertical)
- explanatory variable goes on x-axis (horizontal)

## Simple Linear Regression

- If a scatterplot suggests a linear relationship between 2 variables, we want to summarize the relationship by drawing a straight line on the plot.
- A *regression line* summarizes the relationship between a response variable and an explanatory variable.
  - Both variables must be quantitative.
- definition: A *regression line* is a straight line that describes how a response variable  $Y$  changes as an explanatory variable  $X$  changes.
  - often used to predict the value of  $Y$  that corresponds to a given value of  $X$ .

- Note: Correlation analysis does not distinguish between response and explanatory variables.
- Example: The admissions director of the University of Iowa wants to guess how successful incoming students are likely to be.
- The high school GPA is part of each incoming student’s record. The admissions director wishes to predict the student’s UI GPA.
- What is the response variable and what is the explanatory variable?

## Recall straight lines

$$y = a + bx$$

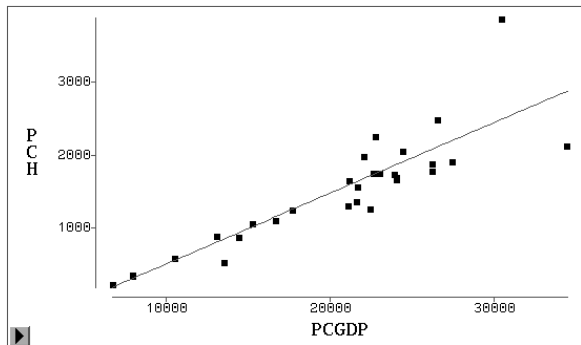
- $a$  : intercept; the value of  $Y$  when  $X = 0$
- $b$  : slope; how much  $Y$  changes when  $X$  increases by 1 unit

## Least squares: choosing the “best” estimated line for a set of sample data

$a$  and  $b$  are estimated by choosing a line as follows:

- for each observed value  $y_i$  in the sample data, compute the distance from  $y_i$  to the line
- square each of the distances
- add up all the squared distances
- choose the line that makes the sum of these squared distances the smallest

- Example: the estimated regression line for the health care expenditures and gross domestic product is



$$\hat{y} = -465.7 + 0.0968x$$

- This means that if country A has 1 unit higher PCGDP than country B, we would expect country A to have 0.0968 higher PCH than country B.

## Using sample data to estimate the intercept and slope

- We will write an estimated regression line based on sample data as

$$\hat{y} = a + bx$$

- $a$  is the estimated intercept, and  $b$  is the estimated slope
- The hat over the  $y$  means that  $\hat{y}$  is the *predicted* value of the response variable, not an actual observed value

- Since we are measuring PCH in dollars and PCGDP in dollars, this means for every additional dollar in PCGDP, we expect about a 9.7-cent increase in PCH.



- Note that it makes no sense in this problem to say that the intercept (-465.7) is the amount of per capital health care expenditure that we would expect in a country with PCGDP = 0.
- **An estimated regression line is meaningful only for the range of X values actually observed.**
  - In the PCH/PCGDP problem, this is about \$8000 - 33000. The estimated intercept makes the linear relationship come out right over this range of X values.