6. Functions of random variables

Readings: Rosner: 5.6
OpenIntro Statistics: 2.4

Homework: Homework 2 due by noon on September 17
Homework 3 due by noon on September 24

Overview

A) Linear functions with constants
B) Sums and differences of independent random variables
C) Difference between expected value E[X] and the sample average $\bar{X}$
D) Sums and difference of dependent random variables
E) Product of independent random variables

Functions of Random Variables

We are often interested in functions of random variables because we need, for example, to change the scale of measurement to something more standard or conventional, or to average a series of measurements, or to take the difference of two observations.

We can derive the properties of these newly created random variables directly from our knowledge of the properties of sums (for discrete random variables) or integrals (for continuous random variables).

A) Linear functions: For constants a and b

$$E(a + bX) = a + b\,E(X) = a + b\mu$$

$$V(a + bX) = V(a) + V(bX) = 0 + b^2V(X) = b^2\sigma^2$$

$$s.d.(a + bX) = |b|\,s.d.(X) = b\sigma$$

Example:

x	0	1	2	3
P(X=x)	0.5	0.3	0.2	0.1

What does the r.v. Y = 2X + 3 look like?

y				
P(Y=y)				

e.g. Suppose a population has mean weight $\mu = E(X) = 70$kg and s.d. = $\sigma = 8$ kg.

What are the mean and s.d. if we change units to lbs. where 1 kg = 2.2 lbs?

$b =$

$E(Y) =$

$V(Y) =$

s.d.$(Y) =$

B) Sums and Differences:

For any r.v. X and Y:

$E(X + Y) = E(X) + E(Y)$

$E(X – Y) = E(X) – E(Y)$

For ***independent*** r.v. X and Y:

$V(X + Y) = V(X) + V(Y)$

$V(X – Y) = V(X) + V(Y)$

s.d.(X + Y) = $\sqrt{V(X) + V(Y)}$ ≠ s.d.(X) + s.d.(Y)

We can use these results to understand sums, differences, averages, etc. used when observing several (*n*) subjects randomly chosen from the same population.

The probability model is X_i = measurement on subject *i*.
If X_i are *iid* (independently, identically distributed) with mean $E(X_i) = \mu$ and $s.d.(X_i) = \sigma$, then:

$$E(X_1 + \cdots + X_n) = E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E(X_i) = E(X_1) + \cdots + E(X_n) = \mu + \cdots + \mu = n\mu$$

$$V(X_1 + \cdots + X_n) = V\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} V(X_i) = V(X_1) + \cdots + V(X_n) = \sigma^2 + \cdots + \sigma^2 = n\sigma^2$$

needs independence for relationship to hold!

From these, we get the ***very important results*** for averages of *iid* random variables $X_1, \ldots, X_n$:

$$E(\bar{X}) = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}E\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}E[X_1 + \cdots + X_n] = \frac{n\mu}{n} = \mu$$

$$V(\bar{X}) = V\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}V\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}V[X_1 + \cdots + X_n] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$s.d.(\bar{X}) = \sqrt{V(\bar{X})} = \frac{\sigma}{\sqrt{n}} = SEM$ (standard error of the mean)

e.g. Let X_i = DBP of subject *i* = 1,…, n - randomly chosen from a population with $E(X_i) = \mu = 80$ mmHg and $s.d.(X_i) = \sigma = 10$ mmHg. Find $E(\bar{X})$ and $s.d.(\bar{X})$ and explain.

$E(\bar{X})$ =

$s.d.(\bar{X})$ =

$s.d.(\bar{X})$ measures the precision with which the sample mean estimates the corresponding population mean, μ. If *n* is large then $s.d.(\bar{X})$ is small. $\bar{X}$ has a smaller s.d. than the X_i, so it is closer to $\mu = 80$ than X_i is, on average.

C) Expected value E[X] vs. sample average $\bar{X}$

Both are weighted averages of the possible values of X

E[X] uses theoretical probabilities

$\bar{X}$ uses sample probabilities (relative frequencies)

Example 1

Suppose you want to estimate the *total* amount of an active drug in 10 pills of a certain kind. Lab #1 proposes to test 10 pills and add the results. Lab #2 proposes to test 1 pill and multiply by 10. Explain using E[X] and V[X] which lab would be best to hire.

Example 2 – comparing means from two samples

Let X_i be the Gestational age of infant *i*, for n = 50 female and n = 50 male infants

$X_{F,i}$ ~ Normal(29, 8); $X_{M,i}$ ~ Normal(29, 4)

Let's consider the linear combination:

$$\bar{X}_F - \bar{X}_M = \left(\frac{1}{50}X_{F,1} + \frac{1}{50}X_{F,2} + \cdots + \frac{1}{50}X_{F,50}\right) - \left(\frac{1}{50}X_{M,1} + \frac{1}{50}X_{M,2} + \cdots + \frac{1}{50}X_{M,50}\right)$$

What is the mean of this linear combination (i.e., what is the mean of $\bar{X}_F - \bar{X}_M$)?

$$E[\bar{X}_F - \bar{X}_M] =$$

What is the variance of this linear combination (i.e., what is the variance of $\bar{X}_F - \bar{X}_M$)?

$$V[\bar{X}_F - \bar{X}_M] =$$

So, $\bar{X}_F - \bar{X}_M \sim N(\_\_\_\_\_, \_\_\_\_\_) = N(\_\_\_\_\_, \_\_\_\_\_)$

The results from Example 2 helped us to obtain the distribution of the difference between two sample means from of a set of observations that are themselves normally distributed with possibly different means and variances.

Knowing this, we will be able to make inferences (based on probability statements) about the differences between two (or more) sample means from normally distributed samples, e.g., using a two-sample t-test, a permutation test based on an underlying t-test, or a bootstrap confidence interval for the difference in two means.

D) Distribution of Sums and Differences when random variables are not independent

Example from Rosner - A hypothesis exists that a high-protein diet may aggravate the course of kidney disease among diabetic patients. To test the feasibility of administering a low-protein diet to such patients, a small "pilot" study is set up where 20 diabetic patients are followed for 1 year on the diet. Serum creatinine is a parameter often used to monitor kidney function.

Let X_1 be the serum creatinine at baseline and X_2 the serum creatinine after 1 year. We wish to compute the expected value and variance of the change in serum creatinine in diabetic patients represented by the random variable $D = X_1 - X_2$, assuming that
$E(X_1) = E(X_2) = \mu = 1.5$ mg/dL and $Var(X_1) = Var(X_2) = \sigma^2 = .25$ $(mg/dL)^2$.

First, we need to ask: Are X_1 and X_2 independent? Why or why not?

The random variables X_1 and X_2 are <u>not</u> independent because they represent serum creatinine values on the <u>same</u> subject over time. The values of serum creatinine will be more similar from day to day, or year to year, on a patient than values between patients.

If that's the case, would you expect the variability within a patient, Var(D), to be larger or smaller than the variability between patients?

What does $Var(X_1) = Var(X_2) = \sigma^2 = .25\ (mg/dL)^2$ above represent, within or between patient variability?

Regardless of independence, we can still compute the expected value of D. Why?

What is E(D) =

What is the Var(D)?

$Var(D) = Var(X_1-X_2) = (1^2)Var(X_1) + (-1^2)Var(X_2)$ +/- _____? How might you expect this to behave?

The covariance between X_1 and X_2 is defined as: $Cov(X_1, X_2) = E[(X_1-\mu)(X_2-\mu)]$.

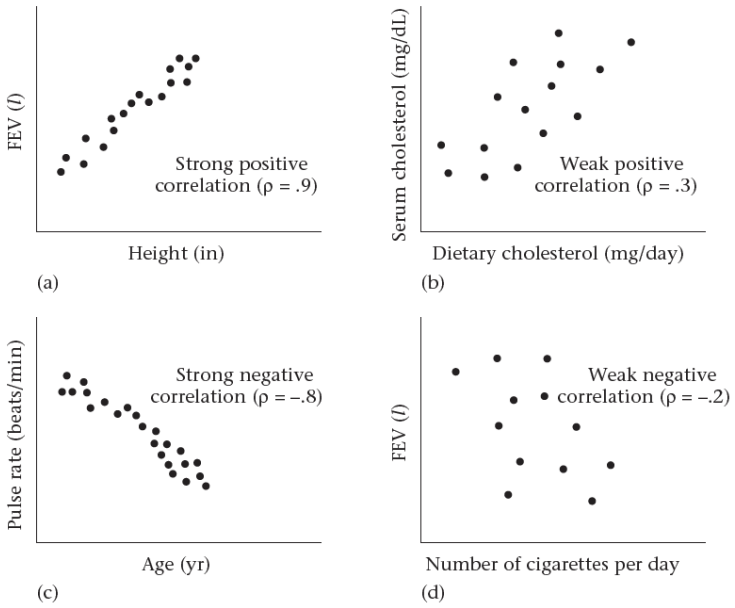
This is related to (Pearson) correlation: $\rho = Corr(X_1, X_2) = Cov(X_1, X_2)/(\sigma_{X1}\sigma_{X2})$.

The Pearson correlation coefficient (ρ) is used to describe the relationship between two continuous variables (e.g., height and weight). It measures the *strength* (qualitatively) and *direction* of the linear relationship between two variables.

A correlation can be between –1 and 1:

- If the correlation is greater than 0, then as X_1 increases, X_2 increases and the two variables are said to be positively correlated. $\rho = 1$ is perfect positive correlation.
- If the correlation is less than 0, then as X_1 increases, X_2 decreases and the two variables are said to be negatively correlated. $\rho = -1$ is perfect negative correlation.
- If the correlation is 0 then there is no linear relationship between X_1 and X_2. The two variables are said to be uncorrelated.

Figure 5.16 Interpretation of various degrees of correlation



Which pattern might we expect two serum creatinine measurements on the same person to follow?

$Var(D) = Var(X_1-X_2) = (1^2)Var(X_1) + (-1^2)\ Var(X_2) + (-1)2*Cov(X_1, X_2)$

Suppose the correlation coefficient between the two determinations of serum creatinine 1 year apart is .5, what's the variance of the change in serum creatinine over 1 year? (In other words, if $\rho = 0.5$, what's $Cov(X_1, X_2)$?)

$0.5 = Cov(X_1, X_2)/\sigma_{X1}\sigma_{X2} = Cov(X_1, X_2)/(0.5*0.5) \Rightarrow Cov(X_1, X_2) = 0.5^3$

$Var(D) = (1^2).25 + (-1^2)(.25) - 2*0.5^3 = .25$

Notice that this variance is much smaller than the variance of the difference in serum creatinine between two different subjects (at the same or different times). In that case, because values for two different subjects are independent, $Corr(X_1, X_2) = 0$, and it follows that $Var(D) = (1^2).25 + (-1^2)(.25) + 0 = .50$

Thus, the rationale for using each person as his or her own control is because it greatly reduces variability. This is relevant to paired-sample vs. independent-sample experimental designs for comparing two groups (such as a treated and a control group).

General result:

$$\boldsymbol{Var}\left[\sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} \boldsymbol{c_i X_i}\right] = \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} \boldsymbol{c_i^2 Var(X_i)} + \boldsymbol{2}\sum_{\boldsymbol{i=1}}^{\boldsymbol{k}}\sum_{\boldsymbol{i<j}}^{\boldsymbol{k}} \boldsymbol{c_i c_j Cov(X_i, X_j)}$$

E) Distribution of Product when X and Y independent

If X, Y are independent, what is E(XY)?

E(XY) = E(X)E(Y) https://en.wikipedia.org/wiki/Product_distribution

What about Var(XY)?

Recall definition: $Var(X) = E(X^2) - E^2(X)$

Var(XY) = (fill-in in class) …

Hospital cost problem – Homework 3

Z = per patient monthly cost ranging from 0 to ∞
R = Bernoulli random variable, 0 or 1 with p = probability of non-zero cost

Let Y = ZR …

Distribution of Product when X and Y not independent - More advanced topic, not covered here.