

Question 1.a.

For which of the variables would the mean **NOT** provide a scientifically meaningful descriptions of the sample? Very briefly explain your reasons (just a few words should suffice to justify your entire answer).

enroll= Date of enrollment in the study (MMDDYY format)
age= Age (years)
sex = Sex (0=male, 1=female)
albumin= Serum albumin (a protein made by the liver and found in the blood) (g/dl)
billi= Serum bilirubin (a breakdown product of hemoglobin that is made by the liver) (mg/dl)
cholest= Serum cholesterol (mg/dl)
edema= Presence of edema (swelling of extremities) (0= no, 1=yes)
sgot = Serum SGOT (an enzyme found in liver cells) (U/l)
spiders = Presence of spider angiomas (superficial tortuous veins seen on the skin) (0= no, 1=yes)
stage= Stage of disease (1= best, 2, 3, or 4= worst)
obstime= Observation time from enrollment until death or until end of study, whichever comes first (years)
status= Survival status at time of last observation time (0=still alive, 1= dead)

November 2, 2011

Question 1.a.

For which of the variables would the mean **NOT** provide a scientifically meaningful descriptions of the sample? Very briefly explain your reasons (just a few words should suffice to justify your entire answer).

enroll= Date of enrollment in the study (MMDDYY format)
age= Age (years)
sex = Sex (0=male, 1=female)
albumin= Serum albumin (a protein made by the liver and found in the blood) (g/dl)
billi= Serum bilirubin (a breakdown product of hemoglobin that is made by the liver) (mg/dl)
cholest= Serum cholesterol (mg/dl)
edema= Presence of edema (swelling of extremities) (0= no, 1=yes)
sgot = Serum SGOT (an enzyme found in liver cells) (U/l)
spiders = Presence of spider angiomas (superficial tortuous veins seen on the skin) (0= no, 1=yes)
stage= Stage of disease (1= best, 2, 3, or 4= worst)
obstime= Observation time from enrollment until death or until end of study, whichever comes first (years)
status= Survival status at time of last observation time (0=still alive, 1= dead)

November 2, 2011

Question 1.b.

For which of the variables would the mean **NOT** be useful when trying to compare distributions across populations. Briefly explain your reasons.

enroll= Date of enrollment in the study (MMDDYY format)
age= Age (years)
sex = Sex (0=male, 1=female)
albumin= Serum albumin (a protein made by the liver and found in the blood) (g/dl)
billi= Serum bilirubin (a breakdown product of hemoglobin that is made by the liver) (mg/dl)
cholest= Serum cholesterol (mg/dl)
edema= Presence of edema (swelling of extremities) (0= no, 1=yes)
sgot = Serum SGOT (an enzyme found in liver cells) (U/l)
spiders = Presence of spider angiomas (superficial tortuous veins seen on the skin) (0= no, 1=yes)
stage= Stage of disease (1= best, 2, 3, or 4= worst)
obstime= Observation time from enrollment until death or until end of study, whichever comes first (years)
status= Survival status at time of last observation time (0=still alive, 1= dead)

November 2, 2011

Question 1.b.

For which of the variables would the mean **NOT** be useful when trying to compare distributions across populations. Briefly explain your reasons.

enroll= Date of enrollment in the study (MMDDYY format)
age= Age (years)
sex = Sex (0=male, 1=female)
albumin= Serum albumin (a protein made by the liver and found in the blood) (g/dl)
billi= Serum bilirubin (a breakdown product of hemoglobin that is made by the liver) (mg/dl)
cholest= Serum cholesterol (mg/dl)
edema= Presence of edema (swelling of extremities) (0= no, 1=yes)
sgot = Serum SGOT (an enzyme found in liver cells) (U/l)
spiders = Presence of spider angiomas (superficial tortuous veins seen on the skin) (0= no, 1=yes)
stage= Stage of disease (1= best, 2, 3, or 4= worst)
obstime= Observation time from enrollment until death or until end of study, whichever comes first (years)
status= Survival status at time of last observation time (0=still alive, 1= dead)

November 2, 2011

Question 2.a.

For which of the variables would the median **NOT** provide a scientifically meaningful descriptions of the sample? Very briefly explain your reasons (just a few words should suffice to justify your entire answer).

enroll= Date of enrollment in the study (MMDDYY format)
age= Age (years)
sex = Sex (0=male, 1=female)
albumin= Serum albumin (a protein made by the liver and found in the blood) (g/dl)
billi= Serum bilirubin (a breakdown product of hemoglobin that is made by the liver) (mg/dl)
cholest= Serum cholesterol (mg/dl)
edema= Presence of edema (swelling of extremities) (0= no, 1=yes)
sgot = Serum SGOT (an enzyme found in liver cells) (U/l)
spiders = Presence of spider angiomas (superficial tortuous veins seen on the skin) (0= no, 1=yes)
stage= Stage of disease (1= best, 2, 3, or 4= worst)
obstime= Observation time from enrollment until death or until end of study, whichever comes first (years)
status= Survival status at time of last observation time (0=still alive, 1= dead)

November 2, 2011

Question 2.a.

For which of the variables would the median **NOT** provide a scientifically meaningful descriptions of the sample? Very briefly explain your reasons (just a few words should suffice to justify your entire answer).

enroll= Date of enrollment in the study (MMDDYY format)
age= Age (years)
sex = Sex (0=male, 1=female)
albumin= Serum albumin (a protein made by the liver and found in the blood) (g/dl)
billi= Serum bilirubin (a breakdown product of hemoglobin that is made by the liver) (mg/dl)
cholest= Serum cholesterol (mg/dl)
edema= Presence of edema (swelling of extremities) (0= no, 1=yes)
sgot = Serum SGOT (an enzyme found in liver cells) (U/l)
spiders = Presence of spider angiomas (superficial tortuous veins seen on the skin) (0= no, 1=yes)
stage= Stage of disease (1= best, 2, 3, or 4= worst)
obstime= Observation time from enrollment until death or until end of study, whichever comes first (years)
status= Survival status at time of last observation time (0=still alive, 1= dead)

November 2, 2011

Question 2.b.

For which of the variables would the median **NOT** be useful when trying to compare distributions across populations. Briefly explain your reasons.

enroll= Date of enrollment in the study (MMDDYY format)
age= Age (years)
sex = Sex (0=male, 1=female)
albumin= Serum albumin (a protein made by the liver and found in the blood) (g/dl)
billi= Serum bilirubin (a breakdown product of hemoglobin that is made by the liver) (mg/dl)
cholest= Serum cholesterol (mg/dl)
edema= Presence of edema (swelling of extremities) (0= no, 1=yes)
sgot = Serum SGOT (an enzyme found in liver cells) (U/l)
spiders = Presence of spider angiomas (superficial tortuous veins seen on the skin) (0= no, 1=yes)
stage= Stage of disease (1= best, 2, 3, or 4= worst)
obstime= Observation time from enrollment until death or until end of study, whichever comes first (years)
status= Survival status at time of last observation time (0=still alive, 1= dead)

November 2, 2011

Question 2.b.

For which of the variables would the median **NOT** be useful when trying to compare distributions across populations. Briefly explain your reasons.

enroll= Date of enrollment in the study (MMDDYY format)
age= Age (years)
sex = Sex (0=male, 1=female)
albumin= Serum albumin (a protein made by the liver and found in the blood) (g/dl)
billi= Serum bilirubin (a breakdown product of hemoglobin that is made by the liver) (mg/dl)
cholest= Serum cholesterol (mg/dl)
edema= Presence of edema (swelling of extremities) (0= no, 1=yes)
sgot = Serum SGOT (an enzyme found in liver cells) (U/l)
spiders = Presence of spider angiomas (superficial tortuous veins seen on the skin) (0= no, 1=yes)
stage= Stage of disease (1= best, 2, 3, or 4= worst)
obstime= Observation time from enrollment until death or until end of study, whichever comes first (years)
status= Survival status at time of last observation time (0=still alive, 1= dead)

November 2, 2011

Question 3.a.

For which of the variables would the standard deviation **NOT** provide a scientifically meaningful descriptions of the sample? Very briefly explain your reasons (just a few words should suffice to justify your entire answer).

enroll= Date of enrollment in the study (MMDDYY format)
age= Age (years)
sex = Sex (0=male, 1=female)
albumin= Serum albumin (a protein made by the liver and found in the blood) (g/dl)
billi= Serum bilirubin (a breakdown product of hemoglobin that is made by the liver) (mg/dl)
cholest= Serum cholesterol (mg/dl)
edema= Presence of edema (swelling of extremities) (0= no, 1=yes)
sgot = Serum SGOT (an enzyme found in liver cells) (U/l)
spiders = Presence of spider angiomas (superficial tortuous veins seen on the skin) (0= no, 1=yes)
stage= Stage of disease (1= best, 2, 3, or 4= worst)
obstime= Observation time from enrollment until death or until end of study, whichever comes first (years)
status= Survival status at time of last observation time (0=still alive, 1= dead)

November 2, 2011

Question 3.a.

For which of the variables would the standard deviation **NOT** provide a scientifically meaningful descriptions of the sample? Very briefly explain your reasons (just a few words should suffice to justify your entire answer).

enroll= Date of enrollment in the study (MMDDYY format)
age= Age (years)
sex = Sex (0=male, 1=female)
albumin= Serum albumin (a protein made by the liver and found in the blood) (g/dl)
billi= Serum bilirubin (a breakdown product of hemoglobin that is made by the liver) (mg/dl)
cholest= Serum cholesterol (mg/dl)
edema= Presence of edema (swelling of extremities) (0= no, 1=yes)
sgot = Serum SGOT (an enzyme found in liver cells) (U/l)
spiders = Presence of spider angiomas (superficial tortuous veins seen on the skin) (0= no, 1=yes)
stage= Stage of disease (1= best, 2, 3, or 4= worst)
obstime= Observation time from enrollment until death or until end of study, whichever comes first (years)
status= Survival status at time of last observation time (0=still alive, 1= dead)

November 2, 2011

Question 3.b.

For which of the variables would the standard deviation **NOT** be useful when trying to compare distributions across populations. Briefly explain your reasons.

enroll= Date of enrollment in the study (MMDDYY format)
age= Age (years)
sex = Sex (0=male, 1=female)
albumin= Serum albumin (a protein made by the liver and found in the blood) (g/dl)
billi= Serum bilirubin (a breakdown product of hemoglobin that is made by the liver) (mg/dl)
cholest= Serum cholesterol (mg/dl)
edema= Presence of edema (swelling of extremities) (0= no, 1=yes)
sgot = Serum SGOT (an enzyme found in liver cells) (U/l)
spiders = Presence of spider angiomas (superficial tortuous veins seen on the skin) (0= no, 1=yes)
stage= Stage of disease (1= best, 2, 3, or 4= worst)
obstime= Observation time from enrollment until death or until end of study, whichever comes first (years)
status= Survival status at time of last observation time (0=still alive, 1= dead)

November 2, 2011

Question 3.b.

For which of the variables would the standard deviation **NOT** be useful when trying to compare distributions across populations. Briefly explain your reasons.

enroll= Date of enrollment in the study (MMDDYY format)
age= Age (years)
sex = Sex (0=male, 1=female)
albumin= Serum albumin (a protein made by the liver and found in the blood) (g/dl)
billi= Serum bilirubin (a breakdown product of hemoglobin that is made by the liver) (mg/dl)
cholest= Serum cholesterol (mg/dl)
edema= Presence of edema (swelling of extremities) (0= no, 1=yes)
sgot = Serum SGOT (an enzyme found in liver cells) (U/l)
spiders = Presence of spider angiomas (superficial tortuous veins seen on the skin) (0= no, 1=yes)
stage= Stage of disease (1= best, 2, 3, or 4= worst)
obstime= Observation time from enrollment until death or until end of study, whichever comes first (years)
status= Survival status at time of last observation time (0=still alive, 1= dead)

November 2, 2011

Question 4.a.

For which of the variables would the minimum and maximum **NOT** provide a scientifically meaningful descriptions of the sample? Very briefly explain your reasons (just a few words should suffice to justify your entire answer).

enroll= Date of enrollment in the study (MMDDYY format)
age= Age (years)
sex = Sex (0=male, 1=female)
albumin= Serum albumin (a protein made by the liver and found in the blood) (g/dl)
billi= Serum bilirubin (a breakdown product of hemoglobin that is made by the liver) (mg/dl)
cholest= Serum cholesterol (mg/dl)
edema= Presence of edema (swelling of extremities) (0= no, 1=yes)
sgot = Serum SGOT (an enzyme found in liver cells) (U/l)
spiders = Presence of spider angiomas (superficial tortuous veins seen on the skin) (0= no, 1=yes)
stage= Stage of disease (1= best, 2, 3, or 4= worst)
obstime= Observation time from enrollment until death or until end of study, whichever comes first (years)
status= Survival status at time of last observation time (0=still alive, 1= dead)

November 2, 2011

Question 4.a.

For which of the variables would the minimum and maximum **NOT** provide a scientifically meaningful descriptions of the sample? Very briefly explain your reasons (just a few words should suffice to justify your entire answer).

enroll= Date of enrollment in the study (MMDDYY format)
age= Age (years)
sex = Sex (0=male, 1=female)
albumin= Serum albumin (a protein made by the liver and found in the blood) (g/dl)
billi= Serum bilirubin (a breakdown product of hemoglobin that is made by the liver) (mg/dl)
cholest= Serum cholesterol (mg/dl)
edema= Presence of edema (swelling of extremities) (0= no, 1=yes)
sgot = Serum SGOT (an enzyme found in liver cells) (U/l)
spiders = Presence of spider angiomas (superficial tortuous veins seen on the skin) (0= no, 1=yes)
stage= Stage of disease (1= best, 2, 3, or 4= worst)
obstime= Observation time from enrollment until death or until end of study, whichever comes first (years)
status= Survival status at time of last observation time (0=still alive, 1= dead)

November 2, 2011

Question 4.b.

For which of the variables would the minimum and maximum **NOT** be useful when trying to compare distributions across populations. Briefly explain your reasons.

enroll= Date of enrollment in the study (MMDDYY format)
age= Age (years)
sex = Sex (0=male, 1=female)
albumin= Serum albumin (a protein made by the liver and found in the blood) (g/dl)
billi= Serum bilirubin (a breakdown product of hemoglobin that is made by the liver) (mg/dl)
cholest= Serum cholesterol (mg/dl)
edema= Presence of edema (swelling of extremities) (0= no, 1=yes)
sgot = Serum SGOT (an enzyme found in liver cells) (U/l)
spiders = Presence of spider angiomas (superficial tortuous veins seen on the skin) (0= no, 1=yes)
stage= Stage of disease (1= best, 2, 3, or 4= worst)
obstime= Observation time from enrollment until death or until end of study, whichever comes first (years)
status= Survival status at time of last observation time (0=still alive, 1= dead)

November 2, 2011

Question 4.b.

For which of the variables would the minimum and maximum **NOT** be useful when trying to compare distributions across populations. Briefly explain your reasons.

enroll= Date of enrollment in the study (MMDDYY format)
age= Age (years)
sex = Sex (0=male, 1=female)
albumin= Serum albumin (a protein made by the liver and found in the blood) (g/dl)
billi= Serum bilirubin (a breakdown product of hemoglobin that is made by the liver) (mg/dl)
cholest= Serum cholesterol (mg/dl)
edema= Presence of edema (swelling of extremities) (0= no, 1=yes)
sgot = Serum SGOT (an enzyme found in liver cells) (U/l)
spiders = Presence of spider angiomas (superficial tortuous veins seen on the skin) (0= no, 1=yes)
stage= Stage of disease (1= best, 2, 3, or 4= worst)
obstime= Observation time from enrollment until death or until end of study, whichever comes first (years)
status= Survival status at time of last observation time (0=still alive, 1= dead)

November 2, 2011

Question 5.a.

For which of the variables would the 25th and 75th percentile **NOT** provide a scientifically meaningful descriptions of the sample? Very briefly explain your reasons (just a few words should suffice to justify your entire answer).

enroll= Date of enrollment in the study (MMDDYY format)
age= Age (years)
sex = Sex (0=male, 1=female)
albumin= Serum albumin (a protein made by the liver and found in the blood) (g/dl)
billi= Serum bilirubin (a breakdown product of hemoglobin that is made by the liver) (mg/dl)
cholest= Serum cholesterol (mg/dl)
edema= Presence of edema (swelling of extremities) (0= no, 1=yes)
sgot = Serum SGOT (an enzyme found in liver cells) (U/l)
spiders = Presence of spider angiomas (superficial tortuous veins seen on the skin) (0= no, 1=yes)
stage= Stage of disease (1= best, 2, 3, or 4= worst)
obstime= Observation time from enrollment until death or until end of study, whichever comes first (years)
status= Survival status at time of last observation time (0=still alive, 1= dead)

November 2, 2011

Question 5.a.

For which of the variables would the 25th and 75th percentile **NOT** provide a scientifically meaningful descriptions of the sample? Very briefly explain your reasons (just a few words should suffice to justify your entire answer).

enroll= Date of enrollment in the study (MMDDYY format)
age= Age (years)
sex = Sex (0=male, 1=female)
albumin= Serum albumin (a protein made by the liver and found in the blood) (g/dl)
billi= Serum bilirubin (a breakdown product of hemoglobin that is made by the liver) (mg/dl)
cholest= Serum cholesterol (mg/dl)
edema= Presence of edema (swelling of extremities) (0= no, 1=yes)
sgot = Serum SGOT (an enzyme found in liver cells) (U/l)
spiders = Presence of spider angiomas (superficial tortuous veins seen on the skin) (0= no, 1=yes)
stage= Stage of disease (1= best, 2, 3, or 4= worst)
obstime= Observation time from enrollment until death or until end of study, whichever comes first (years)
status= Survival status at time of last observation time (0=still alive, 1= dead)

November 2, 2011

Question 5.b.

For which of the variables would the 25th and 75th percentile **NOT** be useful when trying to compare distributions across populations. Briefly explain your reasons.

enroll= Date of enrollment in the study (MMDDYY format)
age= Age (years)
sex = Sex (0=male, 1=female)
albumin= Serum albumin (a protein made by the liver and found in the blood) (g/dl)
billi= Serum bilirubin (a breakdown product of hemoglobin that is made by the liver) (mg/dl)
cholest= Serum cholesterol (mg/dl)
edema= Presence of edema (swelling of extremities) (0= no, 1=yes)
sgot = Serum SGOT (an enzyme found in liver cells) (U/l)
spiders = Presence of spider angiomas (superficial tortuous veins seen on the skin) (0= no, 1=yes)
stage= Stage of disease (1= best, 2, 3, or 4= worst)
obstime= Observation time from enrollment until death or until end of study, whichever comes first (years)
status= Survival status at time of last observation time (0=still alive, 1= dead)

November 2, 2011

Question 5.b.

For which of the variables would the 25th and 75th percentile **NOT** be useful when trying to compare distributions across populations. Briefly explain your reasons.

enroll= Date of enrollment in the study (MMDDYY format)
age= Age (years)
sex = Sex (0=male, 1=female)
albumin= Serum albumin (a protein made by the liver and found in the blood) (g/dl)
billi= Serum bilirubin (a breakdown product of hemoglobin that is made by the liver) (mg/dl)
cholest= Serum cholesterol (mg/dl)
edema= Presence of edema (swelling of extremities) (0= no, 1=yes)
sgot = Serum SGOT (an enzyme found in liver cells) (U/l)
spiders = Presence of spider angiomas (superficial tortuous veins seen on the skin) (0= no, 1=yes)
stage= Stage of disease (1= best, 2, 3, or 4= worst)
obstime= Observation time from enrollment until death or until end of study, whichever comes first (years)
status= Survival status at time of last observation time (0=still alive, 1= dead)

November 2, 2011

Question 6

Where relevant, indicate which of the above variables appear to have a skewed distribution due to outlying values. Briefly explain your reasons.

Variable	N	Mean	SD	Min	25%ile	Median	75%ile	Max
enroll	418	68,475	35,275	10,192	32,791	70,891	100,791	123,090
age	418	50.7	10.4	26.3	42.8	51	58.3	78.4
sex	312	0.885	0.32	0	1	1	1	1
albumin	418	3.5	0.42	1.96	3.24	3.53	3.77	4.64
bili	418	3.22	4.41	0.3	0.8	1.4	3.4	28
cholest	284	370	232	120	249	310	400	1775
edema	418	0.12	0.325	0	0	0	0	1
sgot	312	122.6	56.7	26.4	80.6	114.7	151.9	457.3
spiders	312	0.288	0.454	0	0	0	1	1
stage	312	3.032	0.878	1	2	3	4	4
obstime	418	5.49	3.08	0.11	3.26	5.04	7.40	12.47
status	418	0.385	0.487	0	0	0	1	1

November 2, 2011

Question 6

Where relevant, indicate which of the above variables appear to have a skewed distribution due to outlying values. Briefly explain your reasons.

Variable	N	Mean	SD	Min	25%ile	Median	75%ile	Max
enroll	418	68,475	35,275	10,192	32,791	70,891	100,791	123,090
age	418	50.7	10.4	26.3	42.8	51	58.3	78.4
sex	312	0.885	0.32	0	1	1	1	1
albumin	418	3.5	0.42	1.96	3.24	3.53	3.77	4.64
bili	418	3.22	4.41	0.3	0.8	1.4	3.4	28
cholest	284	370	232	120	249	310	400	1775
edema	418	0.12	0.325	0	0	0	0	1
sgot	312	122.6	56.7	26.4	80.6	114.7	151.9	457.3
spiders	312	0.288	0.454	0	0	0	1	1
stage	312	3.032	0.878	1	2	3	4	4
obstime	418	5.49	3.08	0.11	3.26	5.04	7.40	12.47
status	418	0.385	0.487	0	0	0	1	1

November 2, 2011

Question 7

- The following table presents descriptive statistics for selected variables according to whether the patient did or did not have spider angiomata present. How would you use these descriptive statistics to assess whether the presence of spider angiomata is associated with shortened survival?

Variable	N	Mean	SD	Min	25 th %ile	Median	75 th %ile	Max
<i>Patients without angiomata present</i>								
age	222	50.4	10.6	28.9	42.6	50.1	57	76.7
sex	222	0.856	0.352	0	1	1	1	1
albumin	222	3.58	0.4	2.1	3.35	3.6	3.85	4.64
bili	222	2.43	3.57	0.3	0.7	1.1	2.3	28
edema	222	0.068	0.252	0	0	0	0	1
obstime	222	6.01	3.01	0.11	3.69	5.92	8.19	12.47
status	222	0.329	0.471	0	0	0	1	1
<i>Patients with spider angiomata present</i>								
age	90	49.2	10.5	26.3	41.6	49.5	56	78.4
sex	90	0.956	0.207	0	1	1	1	1
albumin	90	3.37	0.42	1.96	3.12	3.42	3.63	4.19
bili	90	5.3	5.84	0.5	1.3	3.2	6.5	24.5
edema	90	0.244	0.432	0	0	0	0	1
obstime	90	4.21	2.87	0.14	2.05	3.94	6.07	12.32
status	90	0.578	0.497	0	0	1	1	1

November 2, 2011

Question 7

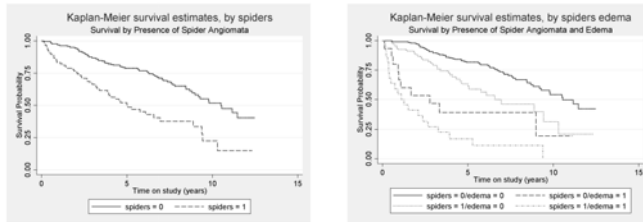
How would you use these descriptive statistics to assess whether the presence of spider angiomata is associated with shortened survival?

- Would not use them because "survival," i.e., time to death here, has censored measurements (is not completely observed on everyone)
- Need Kaplan Meier estimates (e.g. KM curve, KM estimates of percentiles of survival distribution, probability of surviving certain # years)
- What do we need to know to determine the maximum time t at which we could appropriately divide subjects into groups defined by whether or not they survived t years?

November 2, 2011

Question 8

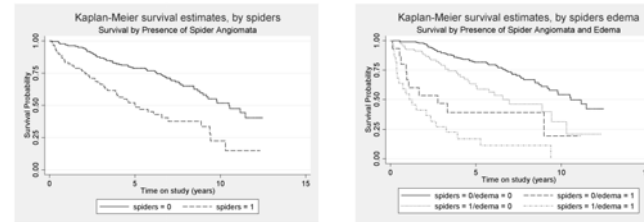
The following are results from a Kaplan-Meier analyses of the time to death within strata defined by the presence of spider angiomata, the presence of edema, or all combinations of those two signs of liver disease.



	Spider Angiomata Absent			Spider Angiomata Present		
	n	2 Year	5 Year	n	2 Year	5 Year
All Patients	211	0.946	0.788	71	0.767	0.520
No Edema	202	0.986	0.817	61	0.882	0.613
Edema	9	0.523	0.379	10	0.419	0.171

Question 8a.

Based on these statistics, would you conclude that there is overall an association between the presence of spider angiomata and the probability of survival? Provide statistics to quantify your answer.



	Spider Angiomata Absent			Spider Angiomata Present		
	n	2 Year	5 Year	n	2 Year	5 Year
All Patients	211	0.946	0.788	71	0.767	0.520
No Edema	202	0.986	0.817	61	0.882	0.613
Edema	9	0.523	0.379	10	0.419	0.171

Question 8a.

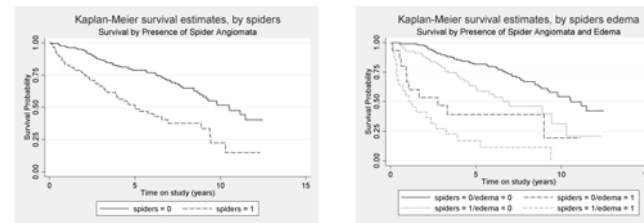
Based on these statistics, would you conclude that there is overall an association between the presence of spider angiomata and the probability of survival? Provide statistics to quantify your answer.

- Patients with spider angiomata tend to survive longer than those without them (5 year survival probabilities of 0.788 and 0.520, respectively, for absolute difference of 0.268).
- Answer the question asked – here you are asked if “overall” there is an association. You are not asked if this association may be due to confounding by another variable or if this association may be modified by another variable. Discussing results within strata would not address the question asked.
- Choose a summary measure to quantify the effect and use descriptive statistics on this summary measure to support your conclusion.

November 2, 2011

Question 8b.

Based on these statistics, would you conclude that the presence of edema modifies any association between the presence of spider angiomata and survival? Provide statistics to quantify your answer.



	Spider Angiomata Absent			Spider Angiomata Present		
	n	2 Year	5 Year	n	2 Year	5 Year
All Patients	211	0.946	0.788	71	0.767	0.520
No Edema	202	0.986	0.817	61	0.882	0.613
Edema	9	0.523	0.379	10	0.419	0.171

Question 8b.

Based on the above statistics, would you conclude that the presence of edema modifies any association between the presence of spider angiomas and survival? Provide statistics to quantify your answer.

	Spider Angiomata Absent			Spider Angiomata Present		
	n	Survival Probability		n	Survival Probability	
		2 Year	5 Year		2 Year	5 Year
All Patients	211	0.946	0.788	71	0.767	0.520
No Edema	202	0.986	0.817	61	0.882	0.613
Edema	9	0.523	0.379	10	0.419	0.171

- One possible answer: Looking at 5 year survival probs, among those without edema, difference is $0.817 - 0.613 = 0.204$ for survival benefit in group without spider angiomas. Among those with edema, difference is $0.379 - 0.171 = 0.208$. These differences are very similar – association between spiders and absolute difference in 5 year survival is similar in those with and without edema – no effect modification.
- Another possible answer: Among those without edema, ratio of 5 year probs is $0.817/0.613 = 1.3$. Among those with edema, this ratio is $0.379/0.171 = 2.3$. These ratios are different – association between spiders and relative difference in 5 year survival is different in those with and without edema – effect modification!

November 2, 2011

Question 8b.

Based on the above statistics, would you conclude that the presence of edema modifies any association between the presence of spider angiomas and survival? Provide statistics to quantify your answer.

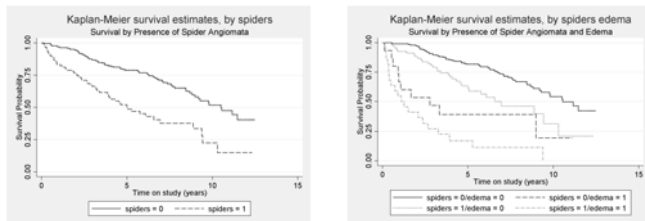
	Spider Angiomata Absent			Spider Angiomata Present		
	n	Survival Probability		n	Survival Probability	
		2 Year	5 Year		2 Year	5 Year
All Patients	211	0.946	0.788	71	0.767	0.520
No Edema	202	0.986	0.817	61	0.882	0.613
Edema	9	0.523	0.379	10	0.419	0.171

- Also could have compared associations between edema and survival in those with and without spider angiomas (effect modification is symmetric). (But to me it's easiest to compare associations of predictor of interest with outcome in strata defined by potential effect modifier)
- Again – also could have chosen different summary measure (conclusion would have been same for 2 year survival)
- Choices of summary measure (2/5 year survival prob, median survival time, etc.) and measure of comparison (absolute difference, ratio, etc.) make a difference in determining effect modification

November 2, 2011

Question 8c.

Based on these statistics, would you conclude that the presence of edema confounds any association between the presence of spider angiomas and survival? Provide statistics to quantify your answer.



	Spider Angiomata Absent			Spider Angiomata Present		
	n	Survival Probability		n	Survival Probability	
		2 Year	5 Year		2 Year	5 Year
All Patients	211	0.946	0.788	71	0.767	0.520
No Edema	202	0.986	0.817	61	0.882	0.613
Edema	9	0.523	0.379	10	0.419	0.171

Question 8c.

Based on these statistics, would you conclude that the presence of edema confounds any association between the presence of spider angiomas and survival? Provide statistics to quantify your answer.

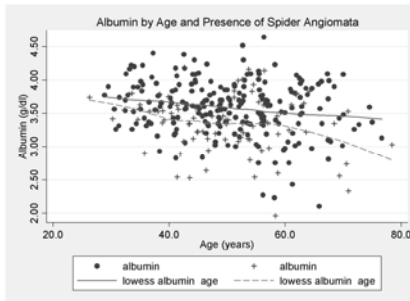
	Spider Angiomata Absent			Spider Angiomata Present		
	n	Survival Probability		n	Survival Probability	
		2 Year	5 Year		2 Year	5 Year
All Patients	211	0.946	0.788	71	0.767	0.520
No Edema	202	0.986	0.817	61	0.882	0.613
Edema	9	0.523	0.379	10	0.419	0.171

- Is edema (potential confounder) associated with spider angiomas (predictor of interest) in the sample?
 - YES: $9/211$ (4%) vs. $10/71$ (14%) of patients without/with spiders have edema
- Is edema (potential confounder) associated with survival (outcome)?
 - YES: difference in 5 year survival probs of $0.817 - 0.379 = 0.438$ and $0.613 - 0.171 = 0.442$ within spider strata
- Alternatively: overall association different from stratum-specific associations
 - difference in 5 year probs = 0.268 overall vs. 0.21 in each strata
 - this only works because the summary measures being compared are means (proportions)

November 2, 2011

Question 9

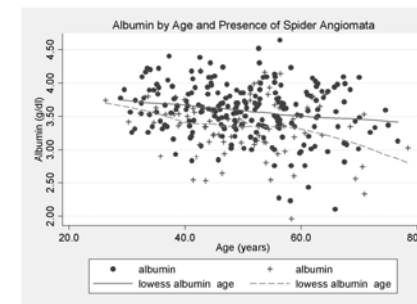
- The following scatter plot displays measurements of serum albumin levels versus age for the sample. Different symbols are used according to whether the patient had spider angiomata (+) or did not have spider angiomata (•). Lowess smooths are superimposed for each stratum defined by the presence of spider angiomata (solid= no angiomata, dashed= spider angiomata present).



November 2, 2011

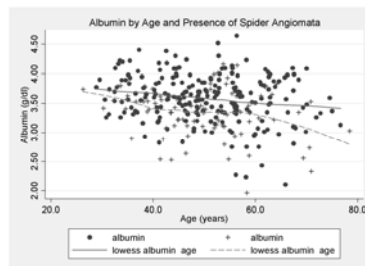
Question 9a.

- What observations would you make about this descriptive analysis?



November 2, 2011

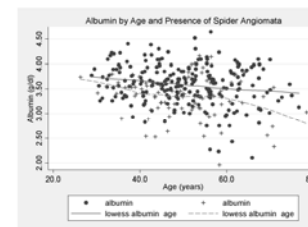
Question 9a.



- Outliers? No obvious ones.
- First order trend? Tendency for lower albumin levels at higher ages.
- Second order? Nah – approximately straight line.
- Within group variability? $\text{Var}(\text{albumin} | \text{age})$ increases with age.

November 2, 2011

Question 9a.

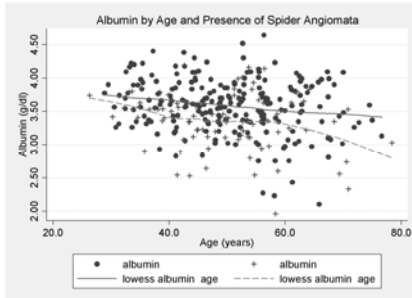


- But we also have presence of spider angiomata on this graph...
- Outliers within spider angiomata strata? No striking ones.
 - Downward linear trend btwn age and albumin in both strata, perhaps slightly steeper slope in those with spider angiomata
 - Trends in within group variance similar in strata
 - Distribution of ages similar in spider angiomata strata
 - Those with spider angiomata tend to have albumin levels compared to those without spiders of same age

November 2, 2011

Question 9b.

- Would you expect the sample correlation between albumin and age to be positive, near zero, or negative in the combined sample?



November 2, 2011

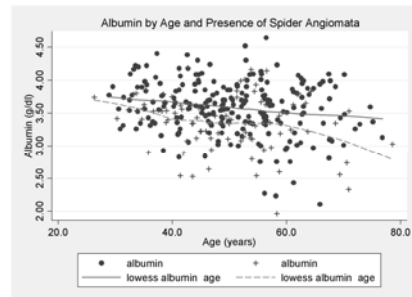
Question 9b.

- Would you expect the sample correlation between albumin and age to be positive, near zero, or negative in the combined sample?
- Correlation should be negative, since best fitting line should have downward slope (as values of one variable are increased, values of other variable tend to decrease)
- $r = -0.18$ here (overall)

November 2, 2011

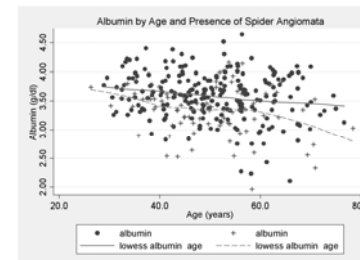
Question 9c.

- If we were to compute the correlations for each stratum separately (i.e., for those with and those without spider angiomata), how do you think they would differ from the correlation in the combined sample? Explain your reasoning.



November 2, 2011

Question 9c.



- Slope of best-fitting lines?
 - Slightly steeper in spider angiomata stratum (would lead to more extreme correlation in that stratum than in combined, more extreme in combined than no spider stratum)
- Variance of predictor (age)?
 - Similar in each stratum and in the combined sample
- Within (age) group variance of albumin?
 - Vertical separation of spider groups leads to greater within group variance in combined: will lead to less extreme correlation in combined sample
- November 2, 2011 $r = 0.23$ in no spiders/spiders strata (vs. -0.18 overall)

Question 10

- Suppose we are interested in studying whether expression of gene DCC can accurately predict the presence of metastases from colon cancer (“metastases” are instances in which the cancer has spread far from its original site). The “gold standard” for the diagnosis of metastases would be based on extensive radiologic examination and surgical exploration. Consider the following study designs for hypothetical studies done at an HMO:
- **Study A:** Using a cancer registry of long term follow-up of patients, we sample 500 patients who are known to have metastatic colon cancer and 500 patients who did not have metastases. We then perform tests for the expression of gene DCC on tumor samples stored from the time of diagnosis.
- **Study B:** We sample 1,000 patients drawn randomly from all colon cancer patients in the registry. Each patient is evaluated for metastases and also has tests performed for the expression of gene DCC on tumor samples stored from the time of diagnosis.
- **Study C:** Using results from previous studies examining the expression of gene DCC in colon cancer patients, we sample 300 patients who had a positive test for DCC expression and 700 patients who had a negative test. We then review the medical records of those patients to assess whether they had metastases

OR
November 2, 2011

Question 10a.

- **Study A:** Using a cancer registry of long term follow-up of patients, we sample 500 patients who are known to have metastatic colon cancer and 500 patients who did not have metastases. We then perform tests for the expression of gene DCC on tumor samples stored from the time of diagnosis.
 - **Study B:** We sample 1,000 patients drawn randomly from all colon cancer patients in the registry. Each patient is evaluated for metastases and also has tests performed for the expression of gene DCC on tumor samples stored from the time of diagnosis.
 - **Study C:** Using results from previous studies examining the expression of gene DCC in colon cancer patients, we sample 300 patients who had a positive test for DCC expression and 700 patients who had a negative test. We then review the medical records of those patients to assess whether they had metastases or not.
- Which of the above study designs can provide an estimate of the prevalence of metastases among all colon cancer patients?

November 2, 2011

Question 10a.

- **Study A:** Using a cancer registry of long term follow-up of patients, we sample 500 patients who are known to have metastatic colon cancer and 500 patients who did not have metastases. We then perform tests for the expression of gene DCC on tumor samples stored from the time of diagnosis.
 - **Study B:** We sample 1,000 patients drawn randomly from all colon cancer patients in the registry. Each patient is evaluated for metastases and also has tests performed for the expression of gene DCC on tumor samples stored from the time of diagnosis.
 - **Study C:** Using results from previous studies examining the expression of gene DCC in colon cancer patients, we sample 300 patients who had a positive test for DCC expression and 700 patients who had a negative test. We then review the medical records of those patients to assess whether they had metastases or not.
- Which of the above study designs can provide an estimate of the prevalence of metastases among all colon cancer patients?
- Study B (cross-sectional design)

November 2, 2011

Question 10b.

- **Study A:** Using a cancer registry of long term follow-up of patients, we sample 500 patients who are known to have metastatic colon cancer and 500 patients who did not have metastases. We then perform tests for the expression of gene DCC on tumor samples stored from the time of diagnosis.
 - **Study B:** We sample 1,000 patients drawn randomly from all colon cancer patients in the registry. Each patient is evaluated for metastases and also has tests performed for the expression of gene DCC on tumor samples stored from the time of diagnosis.
 - **Study C:** Using results from previous studies examining the expression of gene DCC in colon cancer patients, we sample 300 patients who had a positive test for DCC expression and 700 patients who had a negative test. We then review the medical records of those patients to assess whether they had metastases or not.
- Which of the above study designs can provide an estimate of the prevalence of positive DCC expression among all colon cancer patients?

November 2, 2011

Question 10b.

- **Study A:** Using a cancer registry of long term follow-up of patients, we sample 500 patients who are known to have metastatic colon cancer and 500 patients who did not have metastases. We then perform tests for the expression of gene DCC on tumor samples stored from the time of diagnosis.
 - **Study B:** We sample 1,000 patients drawn randomly from all colon cancer patients in the registry. Each patient is evaluated for metastases and also has tests performed for the expression of gene DCC on tumor samples stored from the time of diagnosis.
 - **Study C:** Using results from previous studies examining the expression of gene DCC in colon cancer patients, we sample 300 patients who had a positive test for DCC expression and 700 patients who had a negative test. We then review the medical records of those patients to assess whether they had metastases or not.
- Which of the above study designs can provide an estimate of the prevalence of positive DCC expression among all colon cancer patients?
- B (cross-sectional design)

November 2, 2011

Question 10c., d.

- **Study A:** Using a cancer registry of long term follow-up of patients, we sample 500 patients who are known to have metastatic colon cancer and 500 patients who did not have metastases. We then perform tests for the expression of gene DCC on tumor samples stored from the time of diagnosis.
 - **Study B:** We sample 1,000 patients drawn randomly from all colon cancer patients in the registry. Each patient is evaluated for metastases and also has tests performed for the expression of gene DCC on tumor samples stored from the time of diagnosis.
 - **Study C:** Using results from previous studies examining the expression of gene DCC in colon cancer patients, we sample 300 patients who had a positive test for DCC expression and 700 patients who had a negative test. We then review the medical records of those patients to assess whether they had metastases or not.
- Which of the above study designs can provide an estimate of the patients having a positive DCC test among the patients with metastatic colon cancer? What is this probability usually called?
 - Which of the above study designs can provide an estimate of the patients having a negative DCC test among the patients without metastatic colon cancer? What is this probability usually called?

November 2, 2011

Question 10c., d.

- **Study A:** Using a cancer registry of long term follow-up of patients, we sample 500 patients who are known to have metastatic colon cancer and 500 patients who did not have metastases. We then perform tests for the expression of gene DCC on tumor samples stored from the time of diagnosis.
 - **Study B:** We sample 1,000 patients drawn randomly from all colon cancer patients in the registry. Each patient is evaluated for metastases and also has tests performed for the expression of gene DCC on tumor samples stored from the time of diagnosis.
 - **Study C:** Using results from previous studies examining the expression of gene DCC in colon cancer patients, we sample 300 patients who had a positive test for DCC expression and 700 patients who had a negative test. We then review the medical records of those patients to assess whether they had metastases or not.
- Which of the above study designs can provide an estimate of the patients having a positive DCC test among the patients with metastatic colon cancer? What is this probability usually called?
 - Which of the above study designs can provide an estimate of the patients having a negative DCC test among the patients without metastatic colon cancer? What is this probability usually called?
- A (sampling based on disease status) or B (cross-sectional design)
 - "Sensitivity" = $\Pr(\text{positive test} \mid \text{disease})$
 - "Specificity" = $\Pr(\text{negative test} \mid \text{no disease})$

November 2, 2011

Question 10e., f.

- **Study A:** Using a cancer registry of long term follow-up of patients, we sample 500 patients who are known to have metastatic colon cancer and 500 patients who did not have metastases. We then perform tests for the expression of gene DCC on tumor samples stored from the time of diagnosis.
 - **Study B:** We sample 1,000 patients drawn randomly from all colon cancer patients in the registry. Each patient is evaluated for metastases and also has tests performed for the expression of gene DCC on tumor samples stored from the time of diagnosis.
 - **Study C:** Using results from previous studies examining the expression of gene DCC in colon cancer patients, we sample 300 patients who had a positive test for DCC expression and 700 patients who had a negative test. We then review the medical records of those patients to assess whether they had metastases or not.
- Suppose we want to estimate what proportion of the DCC positive patients will actually have metastatic colon cancer. Which study designs can provide such an estimate? What is this probability usually called?
 - Suppose we want to estimate what proportion of the DCC negative patients will actually be free of metastases. Which study designs can provide such an estimate? What is this probability usually called?

November 2, 2011

Question 10e., f.

- **Study A:** Using a cancer registry of long term follow-up of patients, we sample 500 patients who are known to have metastatic colon cancer and 500 patients who did not have metastases. We then perform tests for the expression of gene DCC on tumor samples stored from the time of diagnosis.
 - **Study B:** We sample 1,000 patients drawn randomly from all colon cancer patients in the registry. Each patient is evaluated for metastases and also has tests performed for the expression of gene DCC on tumor samples stored from the time of diagnosis.
 - **Study C:** Using results from previous studies examining the expression of gene DCC in colon cancer patients, we sample 300 patients who had a positive test for DCC expression and 700 patients who had a negative test. We then review the medical records of those patients to assess whether they had metastases or not.
- Suppose we want to estimate what proportion of the DCC positive patients will actually have metastatic colon cancer. Which study designs can provide such an estimate? What is this probability usually called?
 - Suppose we want to estimate what proportion of the DCC negative patients will actually be free of metastases. Which study designs can provide such an estimate? What is this probability usually called?
 - B (cross-sectional) or C (sampling based on test result)
 - "Positive predictive value" (PPV) = $\Pr(\text{disease} \mid \text{positive test})$ (or PV+)
 - "Negative predictive value" (NPV) = $\Pr(\text{no disease} \mid \text{negative test})$ (or PV-)

November 2, 2011

Question 10g.

- **Study A:** Using a cancer registry of long term follow-up of patients, we sample 500 patients who are known to have metastatic colon cancer and 500 patients who did not have metastases. We then perform tests for the expression of gene DCC on tumor samples stored from the time of diagnosis.
 - **Study B:** We sample 1,000 patients drawn randomly from all colon cancer patients in the registry. Each patient is evaluated for metastases and also has tests performed for the expression of gene DCC on tumor samples stored from the time of diagnosis.
 - **Study C:** Using results from previous studies examining the expression of gene DCC in colon cancer patients, we sample 300 patients who had a positive test for DCC expression and 700 patients who had a negative test. We then review the medical records of those patients to assess whether they had metastases or not.
- Which of the above study designs can provide information regarding an association between a positive DCC test and presence of metastases? Justify your answer.

November 2, 2011

Question 10g.

- **Study A:** Using a cancer registry of long term follow-up of patients, we sample 500 patients who are known to have metastatic colon cancer and 500 patients who did not have metastases. We then perform tests for the expression of gene DCC on tumor samples stored from the time of diagnosis.
 - **Study B:** We sample 1,000 patients drawn randomly from all colon cancer patients in the registry. Each patient is evaluated for metastases and also has tests performed for the expression of gene DCC on tumor samples stored from the time of diagnosis.
 - **Study C:** Using results from previous studies examining the expression of gene DCC in colon cancer patients, we sample 300 patients who had a positive test for DCC expression and 700 patients who had a negative test. We then review the medical records of those patients to assess whether they had metastases or not.
- Which of the above study designs can provide information regarding an association between a positive DCC test and presence of metastases? Justify your answer.
 - Any of these studies can be used to detect an association because an association could be defined by:
 - Sensitivity \neq 1 – Specificity
 - Specificity \neq 1 – Sensitivity
 - PPV \neq 1 – NPV
 - NPV \neq 1 – PPV

November 2, 2011

Question 10h.

- **Study A:** Using a cancer registry of long term follow-up of patients, we sample 500 patients who are known to have metastatic colon cancer and 500 patients who did not have metastases. We then perform tests for the expression of gene DCC on tumor samples stored from the time of diagnosis.
 - **Study B:** We sample 1,000 patients drawn randomly from all colon cancer patients in the registry. Each patient is evaluated for metastases and also has tests performed for the expression of gene DCC on tumor samples stored from the time of diagnosis.
 - **Study C:** Using results from previous studies examining the expression of gene DCC in colon cancer patients, we sample 300 patients who had a positive test for DCC expression and 700 patients who had a negative test. We then review the medical records of those patients to assess whether they had metastases or not.
- Which of the above study designs would be the easiest to perform logistically?

November 2, 2011

Question 10h.

- **Study A:** Using a cancer registry of long term follow-up of patients, we sample 500 patients who are known to have metastatic colon cancer and 500 patients who did not have metastases. We then perform tests for the expression of gene DCC on tumor samples stored from the time of diagnosis.
 - **Study B:** We sample 1,000 patients drawn randomly from all colon cancer patients in the registry. Each patient is evaluated for metastases and also has tests performed for the expression of gene DCC on tumor samples stored from the time of diagnosis.
 - **Study C:** Using results from previous studies examining the expression of gene DCC in colon cancer patients, we sample 300 patients who had a positive test for DCC expression and 700 patients who had a negative test. We then review the medical records of those patients to assess whether they had metastases or not.
- Which of the above study designs would be the easiest to perform logistically?
- Probably Study A (case-control) study
 - Could identify cases easily for this rare disease
 - Studies B and C: probably wouldn't have enough metastases unless sample size (and # genetic tests) increased
 - Generally case-control easiest with rare disease, cohort easiest with rare exposure

November 2, 2011