Class 26: review for final exam
18.05, Spring 2014

Probability

- **Counting**
  - Sets
  - Inclusion-exclusion principle
  - Rule of product (multiplication rule)
  - Permutation and combinations
- **Basics**
  - Outcome, sample space, event
  - Discrete, continuous
  - Probability function
  - Conditional probability
  - Independent events
  - Law of total probability
  - Bayes' theorem
- **Random variables**
  - Discrete: general, uniform, Bernoulli, binomial, geometric
  - Continuous: general, uniform, normal, exponential
  - pmf, pdf, cdf
  - Expectation = mean = average value
  - Variance; standard deviation
- **Joint distributions**
  - Joint pmf and pdf
  - Independent random variables
  - Covariance and correlation
- **Central limit theorem**

Statistics

- **Maximum likelihood**
- **Least squares**
- **Bayesian inference**
  - Discrete sets of hypotheses
  - Continuous ranges of hypotheses
  - Beta distributions
  - Conjugate priors
  - Choosing priors
  - Probability intervals
- **Frequentist inference**
  - NHST: rejection regions, significance
  - NHST: p-values
  - z, t, χ^2
  - NHST: type I and type II error
  - NHST: power
  - Confidence intervals
- **Bootstrap confidence intervals**

  - Empirical bootstrap confidence intervals
  - Parametric bootstrap confidence intervals
- **Linear regression**

Sets and counting

- Sets: $\emptyset$, union, intersection, complement Venn diagrams, products
- Counting: inclusion-exclusion, rule of product, permutations ${}_nP_k$, combinations ${}_nC_k = \binom{n}{k}$

Problem 1. Consider the nucleotides A, G, C, T.

(a) How many ways are there to make a sequence of 5 nucleotides.

(b) How many sequences of length 5 are there where no adjacent nucleotides are the same

(c) How many sequences of length 5 have exactly one A?

Problem 2. **(a)** How many 5 card poker hands are there?

(b) How many ways are there to get a full house (3 of one rank and 2 of another)?

(c) What's the probability of getting a full house?

Problem 3. **(a)** How many arrangements of the letters in the word probability are there?

(b) Suppose all of these arrangements are written in a list and one is chosen at random. What is the probability it begins with 'b'.

Probability

- Sample space, outcome, event, probability function. Rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
  Special case: $P(A^c) = 1 - P(A)$
  (A and B disjoint $\Rightarrow P(A \cup B) = P(A) + P(B)$.)
- Conditional probability, multiplication rule, trees, law of total probability, independence
- Bayes' theorem, base rate fallacy

Problem 4. Let E and F be two events. Suppose the probability that at least one of them occurs is 2/3. What is the probability that neither E nor F occurs?

Problem 5. Let C and D be two events with $P(C) = 0.3$, $P(D) = 0.4$, and $P(C^c \cap D) = 0.2$.

What is $P(C \cap D)$?

Problem 6. Suppose we have 8 teams labeled T_1, ..., T_8. Suppose they are ordered by placing their names in a hat and drawing the names out one at a time.

(a) How many ways can it happen that all the odd numbered teams are in the odd numbered slots and all the even numbered teams are in the even numbered slots?

(b) What is the probability of this happening?

Problem 7. Suppose you want to divide a 52 card deck into four hands with 13 cards each. What is the probability that each hand has a king?

Problem 8. Suppose we roll a fair die twice. Let A be the event 'the sum of the rolls is 5' and let B be the event 'at least one of the rolls is 4.'

(a) Calculate $P(A|B)$.

(b) Are A and B independent?

Problem 9. On a quiz show the contestant is given a multiple choice question with 4 options. Suppose there is a 70% chance the contestant actually knows the answer. If they don't know the answer they guess with a 25% chance of getting it right. Suppose they get it right. What is the probability that they were guessing?

Problem 10. Suppose you have an urn containing 7 red and 3 blue balls. You draw three balls at random. On each draw, if the ball is red you set it aside and if the ball is blue you put it back in the urn. What is the probability that the third draw is blue?

(If you get a blue ball it counts as a draw even though you put it back in the urn.)

Problem 11. Independence

Suppose that $P(A) = 0.4, P(B) = 0.3$ and $P((A \cup B)^C) = 0.42$. Are A and B independent?

Problem 12. Suppose that events A, B and C are *mutually independent* with

$$P(A) = 0.3, \quad P(B) = 0.4, \quad P(C) = 0.5.$$

Compute the following: (Hint: Use a Venn diagram)

(i) $P(A \cap B \cap C^c)$ (ii) $P(A \cap B^c \cap C)$ (iii) $P(A^c \cap B \cap C)$

Problem 13. Suppose A and B are events with $0 < P(A) < 1$ and $0 < P(B) < 1$.

(a) If A and B are disjoint, can they be independent?

(b) If A and B are independent, can they be disjoint?

Random variables, expectation and variance

- Discrete random variables: events, pmf, cdf

- Bernoulli(p), binomial(n, p), geometric(p), uniform(n)
- $E(X)$, meaning, algebraic properties, $E(h(X))$
- Var(X), meaning, algebraic properties
- Continuous random variables: pdf, cdf
- uniform(a,b), exponential(λ), normal(μ,σ)
- Transforming random variables
- Quantiles

Problem 14. Directly from the definitions of expected value and variance, compute $E(X)$ and Var(X) when X has probability mass function given by the following table:

X	-2	-1	0	1	2
p(X)	1/15	2/15	3/15	4/15	5/15

Problem 15. (Expected value and variance)
Suppose that X takes values between 0 and 1 and has probability density function $2x$. Compute Var(X) and Var(X^2).

Problem 16. The pmf of X is given by the following table

Value of X	-1	0	1
Probability	1/3	1/6	1/2

(a) Compute $E(X)$.
(b) Give the pdf of $Y = X^2$ and use it to compute $E(Y)$.
(c) Instead, compute $E(X^2)$ directly from an extended table.
(d) Compute Var(X).

Problem 17. Compute the expectation and variance of a Bernoulli(p) random variable.

Problem 18. Suppose 100 people all toss a hat into a box and then proceed to randomly pick out a hat. What is the expected number of people to get their own hat back.

Hint: express the number of people who get their own hat as a sum of random variables whose expected value is easy to compute.

pmf, pdf, cdf

Probability Mass Functions, Probability Density Functions and Cumulative Distribution Functions

Problem 19. Suppose that $X \sim \text{Bin}(n, 0.5)$. Find the probability mass function of $Y = 2X$.

Problem 20. Suppose that X is uniform on $[0,1]$. Compute the pdf and cdf of X. If $Y = 2X + 5$, compute the pdf and cdf of Y.

Problem 21. Now suppose that X has probability density function $f_X(x) = \lambda e^{-\lambda x}$ for $x \geq 0$. Compute the cdf, $F_X(x)$. If $Y = X^2$, compute the pdf and cdf of Y.

Problem 22. Suppose that X is a random variable that takes on values 0, 2 and 3 with probabilities 0.3, 0.1, 0.6 respectively. Let $Y = 3(X-1)^2$.

(a) What is the expectation of X?

(b) What is the variance of X?

(c) What is the expection of Y?

(d) Let $F_Y(t)$ be the cumulative density function of Y. What is $F_Y(7)$?

Problem 23. Suppose you roll a fair 6-sided die 25 times (independently), and you get \$3 every time you roll a 6.

Let X be the total number of dollars you win.

(a) What is the pmf of X.

(b) Find $E(X)$ and $\mathrm{Var}(X)$.

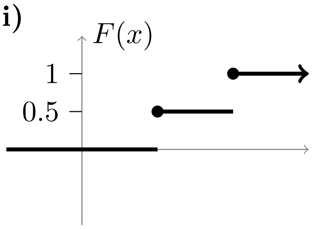
(c) Let Y be the total won on another 25 independent rolls. Compute and compare $E(X+Y)$, $E(2X)$, $\mathrm{Var}(X+Y)$, $\mathrm{Var}(2X)$.

Explain briefly why this makes sense.

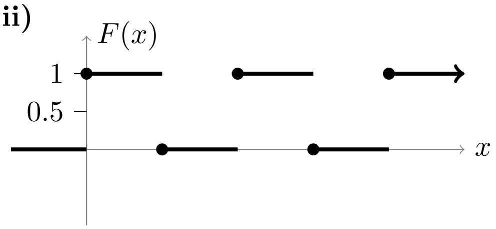
Problem 24. (Continuous random variables)
A continuous random variable X has PDF $f(x) = x + ax^2$ on [0,1]
Find a, the CDF and $P(.5 < X < 1)$.

Problem 25. For each of the following say whether it can be the graph of a cdf. If it can be, say whether the variable is discrete or continuous.

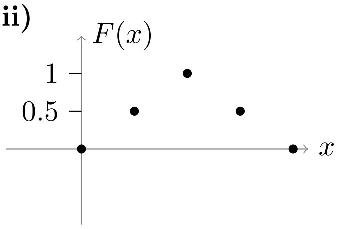
(i)



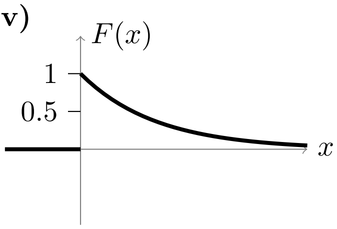
(ii)



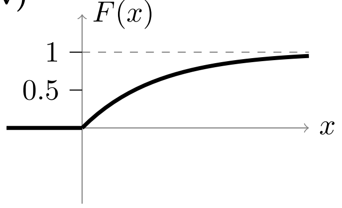
(iii)



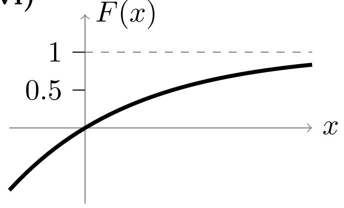
(iv)



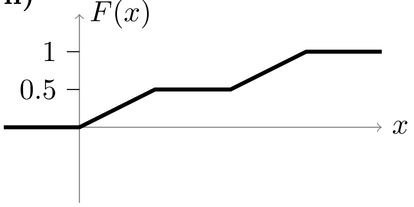
(v)



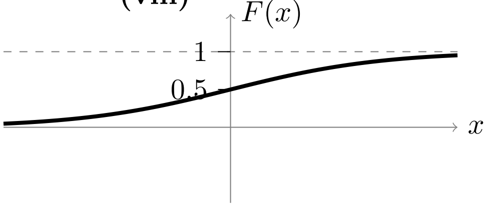
(vi)



(vii)



(viii)



Distributions with names

Problem 26. (Exponential distribution)
Suppose that buses arrive are scheduled to arrive at a bus stop at noon but are always X minutes late, where X is an exponential random variable with probability density function $f_X(x) = \lambda e^{-\lambda x}$. Suppose that you arrive at the bus stop precisely at noon.

(a) Compute the probability that you have to wait for more than five minutes for the bus to arrive.

(b) Suppose that you have already waiting for 10 minutes. Compute the probability that you have to wait an additional five minutes or more.

Problem 27. Normal Distribution: Throughout these problems, let ϕ and Φ be the pdf and cdf, respectively, of the standard normal distribution Suppose Z is a standard normal random variable and let $X = 3Z + 1$.

(a) Express $P(X \leq x)$ in terms of Φ

(b) Differentiate the expression from (a) with respect to x to get the pdf of X, $f(x)$. Remember that $\Phi'(z) = \phi(z)$ and don't forget the chain rule

(c) Find $P(-1 \leq X \leq 1)$

(d) Recall that the probability that Z is within one standard deviation of its mean is approximately 68%. What is the probability that X is within one standard deviation of its mean?

Problem 28. Transforming Normal Distributions
Suppose $Z \sim$ N(0,1) and $Y = e^Z$.

(a) Find the cdf $F_Y(a)$ and pdf $f_Y(y)$ for Y. (For the CDF, the best you can do is write it in terms of Φ the standard normal cdf.)

(b) We don't have a formula for $\Phi(z)$ so we don't have a formula for quantiles. So we have to write quantiles *in terms* of Φ^{-1}.
(i) Write the .33 quantile of Z in terms of Φ^{-1}

(ii) Write the .9 quatntile of Y in terms of Φ^{-1}.
(iii) Find the median of Y.

Problem 29. (Random variables derived from normal r.v.)

Let X_1, X_2, $\dots X_n$ be i.i.d. N$(0,1)$ random variables.
Let $Y_n = X_1^2 + \ldots + X_n^2$.

(a) Use the formula $\text{Var}(X_j) = E(X_j^2) - E(X_j)^2$ to show $E(X_j^2) = 1$.

(b) Set up an integral in x for computing $E(X_j^4)$.

For 3 extra credit points, use integration by parts show $E(X_j^4) = 3$.
(If you don't do this, you can still use the result in part c.)

(c) Deduce from parts (a) and (b) that $\text{Var}(X_j^2) = 2$.

(d) Use the Central Limit Theorem to approximate $P(Y_{100} > 110)$.

Problem 30. More Transforming Normal Distributions

(a) Suppose Z is a standard normal random variable and let $Y = aZ + b$, where $a > 0$ and b are constants.

Show $Y \sim \text{N}(b, a^2)$.

(b) Suppose $Y \sim \text{N}(\mu, \sigma^2)$. Show $\dfrac{Y - \mu}{\sigma}$ follows a standard normal distribution.

Problem 31. (Sums of normal random variables)

Let X be independent random variables where $X \sim N(2, 5)$ and $Y \sim N(5, 9)$ (we use the notation $N(\mu, \sigma^2)$). Let $W = 3X - 2Y + 1$.

(a) Compute $E(W)$ and $\text{Var}(W)$.

(b) It is known that the sum of independent normal distributions is normal. Compute $P(W \le 6)$.

Problem 32. Let $X \sim \text{U}(a, b)$. Compute $E(X)$ and $\text{Var}(X)$.

Problem 33. In $n+m$ independent Bernoulli(p) trials, let S_n be the number of successes in the first n trials and T_m the number of successes in the last m trials.

(a) What is the distribution of S_n? Why?

(b) What is the distribution of T_m? Why?

(c) What is the distribution of $S_n + T_m$? Why?

(d) Are S_n and T_m independent? Why?

Problem 34. Compute the median for the exponential distribution with parameter λ.

Joint distributions

- Joint pmf, pdf, cdf.
- Marginal pmf, pdf, cdf
- Covariance and correlation.

Problem 35. Correlation

Flip a coin 3 times. Use a joint pmf table to compute the covariance and correlation between the number of heads on the first 2 and the number of heads on the last 2 flips.

Problem 36. Correlation

Flip a coin 5 times. Use properties of covariance to compute the covariance and correlation between the number of heads on the first 3 and last 3 flips.

Problem 37. Toss a fair coin 3 times. Let X = the number of heads on the first toss, Y the total number of heads on the last two tosses, and Z the number of heads on the first two tosses.

(a) Give the joint probability table for X and Y. Compute $\text{Cov}(X, Y)$.

(b) Give the joint probability table for X and Z. Compute $\text{Cov}(X, Z)$.

Problem 38. Let X be a random variable that takes values -2, -1, 0, 1, 2; each with probability 1/5. Let $Y = X^2$.

(a) Fill out the following table giving the joint frequency function for X and Y. Be sure to include the marginal probabilities.

X	-2	-1	0	1	2	total
Y						
0						
1						
4						
total						

(b) Find $E(X)$ and $E(Y)$.

(c) Show X and Y are not independent.

(d) Show $\text{Cov}(X, Y) = 0$.

This is an example of uncorrelated but non-independent random variables. The reason this can happen is that correlation only measures the linear dependence between the two variables. In this case, X and Y are not at all linearly related.

Problem 39. Continuous Joint Distributions

Suppose X and Y are continuous random variables with joint density function $f(x, y) = x+y$ on the unit square $[0, 1] \times [0, 1]$.

(a) Let $F(x, y)$ be the joint CDF. Compute $F(1, 1)$. Compute $F(x, y)$.

(b) Compute the marginal densities for X and Y.

(c) Are X and Y independent?

(d) Compute $E(X)$, (Y), $E(X^2+Y^2)$, $\text{Cov}(X,Y)$.

Law of Large Numbers, Central Limit Theorem

Problem 40. Suppose $X_1,\ldots,X_{100}$ are i.i.d. with mean 1/5 and variance 1/9. Use the central limit theorem to estimate $P(\sum X_i < 30)$.

Problem 41. All or None
You have \$100 and, never mind why, you must convert it to \$1000. Anything less is no good. Your only way to make money is to gamble for it. Your chance of winning one bet is p.
Here are two extreme strategies:

Maximum strategy: bet as much as you can each time. To be smart, if you have less than \$500 you bet it all. If you have more, you bet enough to get to \$1000.

Minimum strategy: bet \$1 each time.

If $p < .5$ (the odds are against you) which is the better strategy?

What about $p > .5$ or $p = .5$?

Problem 42. (More Central Limit Theorem)
The average IQ in a population is 100 with standard deviation 15 (by definition, IQ is normalized so this is the case). What is the probability that a randomly selected group of 100 people has an average IQ above 115?

Problem 43. Hospitals (binomial, CLT, etc)

- A certain town is served by two hospitals.
- Larger hospital: about 45 babies born each day.
- Smaller hospital about 15 babies born each day.
- For a period of 1 year, each hospital recorded the days on which more than 60% of the babies born were boys.

(a) Which hospital do you think recorded more such days?

(i) The larger hospital. (ii) The smaller hospital.
(iii) About the same (that is, within 5% of each other).

(b) Let L_i (resp., S_i) be the Bernoulli random variable which takes the value 1 if more than 60% of the babies born in the larger (resp., smaller) hospital on the i^{th} day were boys. Determine the distribution of L_i and of S_i.

(c) Let L (resp., S) be the number of days on which more than 60% of the babies born in the larger (resp., smaller) hospital were boys. What type of distribution do L and S have? Compute the expected value and variance in each case.

(d) Via the CLT, approximate the .84 quantile of L (resp., S). Would you like to revise your answer to part (a)?

(e) What is the correlation of L and S? What is the joint pmf of L and S? Visualize the region corresponding to the event $L > S$. Express $P(L > S)$ as a double sum.

Post unit 2:

1. Confidence intervals
2. Bootstrap confidence intervals
3. Linear regression

Problem 44. Confidence interval 1: basketball

Suppose that against a certain opponent the number of points the MIT basketaball team scores is normally distributed with unknown mean θ and unknown variance, σ^2.

Suppose that over the course of the last 10 games between the two teams MIT scored the following points:

$$59, 62, 59, 74, 70, 61, 62, 66, 62, 75$$

Compute a 95% t–confidence interval for θ. Does 95% confidence mean that the probability θ is in the interval you just found is 95%?

Problem 45. Confidence interval 2

The volume in a set of wine bottles is known to follow a N$(\mu, 25)$ distribution. You take a sample of the bottles and measure their volumes. How many bottles do you have to sample to have a 95% confidence interval for μ with width 1?

Problem 46. Polling confidence intervals

You do a poll to see what fraction p of the population supports candidate A over candidate B. How many people do you need to poll to know p to within 1% with 95% confidence?

Problem 47. Polling confidence intervals 2

Let p be the fraction of the population who prefer candidate A. If you poll 400 people, how many have to prefer candidate A to make the 90% confidence interval entirely in the range where A is preferred.

Problem 48. Confidence intervals 3

Suppose you made 40 confidence intervals with confidence level 95%. About how many of them would you expect to be "wrong'? That is, how many would not actually contain the parameter being estimated? Should you be surprised if 10 of them are wrong?

Problem 49. (Confidence intervals)
A statistician chooses 20 randomly selected class days and counts the number of students present in 18.05. The find a standard deviation of 4.56 students If the number of students present is normally distributed, find the 95% confidence interval for the population standard deviation of the number of students in attendance.

Problem 50. Linear regression (least squares)

(a) Set up fitting the least squares line through the points $(1, 1)$, $(2, 1)$, and $(3, 3)$.

Also see the exam 2 and post exam 2 practice material and the practice final.

MIT OpenCourseWare
https://ocw.mit.edu

18.05 Introduction to Probability and Statistics
Spring 2014

For information about citing these materials or our Terms of Use, visit: https://ocw.mit.edu/terms.