Review for Final Exam
18.05 Spring 2014

Jeremy Orloff and Jonathan Bloom

THANK YOU !!!!

JON !!

PETER !!

RUTHI !!

ERIKA !!

ALL OF YOU !!!!

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

Probability

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

Bootstrapping

Problem **17**.

Directly from the definitions of expected value and variance, compute $E(X)$ and $\mathrm{Var}(X)$ 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 **18**.

Suppose that X takes values between 0 and 1 and has probability density function $2x$. Compute $\mathrm{Var}(X)$ and $\mathrm{Var}(X^2)$.

Problem **20**.

For a certain random variable X it is known that $E(X) = 2$ and $\mathrm{Var}(X) = 3$. What is $E(X^2)$?

Problem **21**.

Determine the expectation and variance of a Bernoulli(p) random variable.

Problem **22**.

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 **27**.

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.

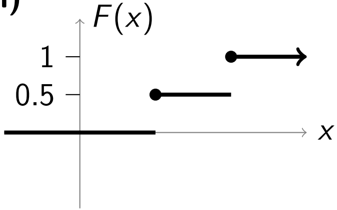
Problem **28**.

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

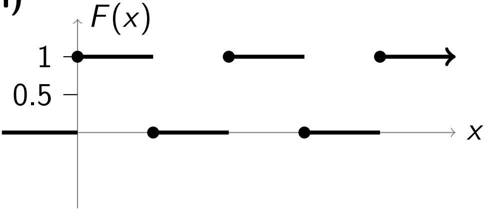
Problem **32**.

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.

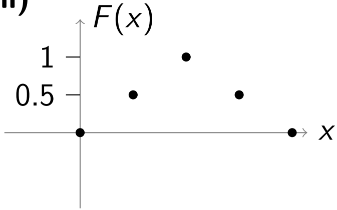
(i)



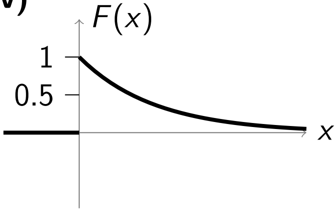
(ii)



(iii)

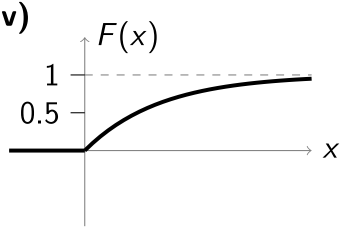


(iv)

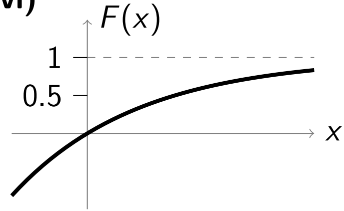


Continued

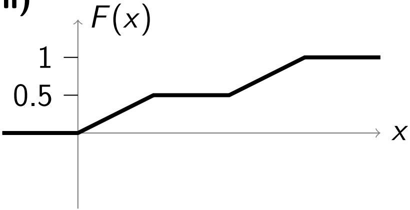
(v)



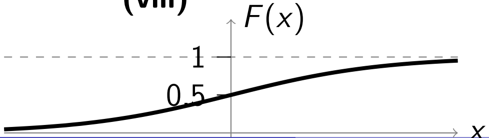
(vi)



(vii)



(viii)



Distributions with names

Problem **35**.

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 **39**.

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

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

Problem **40**.

(**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 $\mathrm{Var}(W)$.
(b) It is known that the sum of independent normal distributions is normal. Compute $P(W \leq 6)$.

Problem **41**.

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

Problem **42**.

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 **43**.

Compute the median for the exponential distribution with parameter λ.

Joint distributions

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

Problem **46**.

To investigate the relationship between hair color and eye color, the hair color and eye color of 5383 persons was recorded. Eye color is coded by the values 1 (Light) and 2 (Dark), and hair color by 1 (Fair/red), 2 (Medium), and 3 (Dark/black). The data are given in the following table:

Eye \ Hair	1	2	3
1	1168	825	305
2	573	1312	1200

The table is turned into a joint pdf for X (hair color) and Y (eye color).

(a) Determine the joint and marginal pmf of X and Y.

(b) Are X and Y independent?

Problem **47**.

Let X and Y be two continuous random variables with joint pdf

$$f(x,y) = \frac{12}{5}xy(1+y) \quad \text{for } 0 \le x \le 1 \text{ and } 0 \le y \le 1,$$

and $f(x) = 0$ otherwise.
(a) Find the probability $P(\frac{1}{4} \le X \le \frac{1}{2}, \frac{1}{3} \le Y \le \frac{2}{3})$.
(b) Determine the joint cdf of X and Y for a and b between 0 and 1.
(c) Use your answer from (b) to find marginal cdf $F_X(a)$ for a between 0 and 1.
(d) Find the marginal pdf $f_X(x)$ directly from $f(x,y)$ and check that it is the derivative of $F_X(x)$.
(e) Are X and Y independent?

Problem **50**.

(**Arithmetic Puzzle**) The joint pmf of X and Y is partly given in the following table.

$X \setminus Y$	0	1	2	
-1	$\dots$	$\dots$	$\dots$	$1/2$
1	$\dots$	$1/2$	$\dots$	$1/2$
	$1/6$	$2/3$	$1/6$	1

(a) Complete the table.
(b) Are X and Y independent?

Problem **51**.

(**Simple Joint Probability**) Let X and Y have joint pmf given by the table:

$X \setminus Y$	1	2	3	4
1	16/136	3/136	2/136	13/136
2	5/136	10/136	11/136	8/136
3	9/136	6/136	7/136	12/136
4	4/136	15/136	14/136	1/136

Compute:
(a) $P(X = Y)$.
(b) $P(X + Y = 5)$.
(c) $P(1 < X \leq 3, 1 < Y \leq 3)$.
(d) $P((X, Y) \in \{1, 4\} \times \{1, 4\})$.

Problem **52**.

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 **54**.

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 **55**.

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 **57**.

(**Central Limit Theorem**)
Let $X_1, X_2, \ldots, X_{144}$ be i.i.d., each with expected value $\mu = E(X_i) = 2$, and variance $\sigma^2 = \text{Var}(X_i) = 4$. Approximate $P(X_1 + X_2 + \cdots X_{144} > 264)$, using the central limit theorem.

Problem **59**.

(**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?

Post unit 2:

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

Confidence intervals 1

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:

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

(a) 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%?

Confidence intervals 1

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:

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

(a) 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%?

answer: Data mean and variance $\bar{x} = 65$, $s^2 = 35.778$. The number of degrees of freedom is 9. We look up $t_{9,.025} = 2.262$ in the t-table The 95% confidence interval is

$$\left[\bar{x} - \frac{t_{9,.025}\, s}{\sqrt{n}},\ \bar{x} + \frac{t_{9,.025}\, s}{\sqrt{n}}\right] = \left[65 - 2.262\sqrt{3.5778},\ 65 + 2.262\sqrt{3.5778}\right]$$

Confidence interval 2

The volume in a set of wine bottles is known to follow a $\mathrm{N}(\mu, 25)$ 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?

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?

answer: Suppose we have taken data $x_1, \ldots, x_n$ with mean $\bar{x}$.
Remember in these probabilities μ is a given (fixed) hypothesis.

$$P(|\bar{x}-\mu| \leq .5 \mid \mu) = .95 \;\Leftrightarrow\; P\left(\frac{|\bar{x}-\mu|}{\sigma/\sqrt{n}} < \frac{.5}{\sigma/\sqrt{n}} \mid \mu\right) = .95 \;\Leftrightarrow\; P$$

Using the table, we have precisely that $\dfrac{.5\sqrt{n}}{5} = 1.96$. So,
$n = (19.6)^2 = \boxed{384.}$.
If we use our rule of thumb that the .95 interval is 2σ we have
$\sqrt{n}/10 = 2 \;\Rightarrow\; n = 400$.

Polling confidence intervals

You do a poll to see what fraction p of the population supports candidate A over candidate B.

1. How many people do you need to poll to know p to within 1% with 95% confidence?

Polling confidence intervals

You do a poll to see what fraction p of the population supports candidate A over candidate B.

1. How many people do you need to poll to know p to within 1% with 95% confidence?

answer: The rule-of-thumb is that a 95% confidence interval is $\bar{x} \pm 1/\sqrt{n}$. To be within 1% we need

$$\frac{1}{\sqrt{n}} = .01 \;\Rightarrow\; n = 10000.$$

Using $z_{.025} = 1.96$ instead the 95% confidence interval is

$$\bar{x} \pm \frac{z_{.025}}{2\sqrt{n}}.$$

To be within 1% we need

$$\frac{z_{.025}}{2\sqrt{n}} = .01 \;\Rightarrow\; n = 9604.$$

Note, we are using the standard Bernoulli approximation $\sigma \leq 1/2$.

Polling confidence intervals 2

2. 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.

Polling confidence intervals 2

2. 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.

answer: The 90% confidence interval is

$$\overline{x} \pm \frac{z_{.05}}{2\sqrt{n}} = \overline{x} \pm \frac{1.64}{40}$$

We want $\overline{x} - \frac{1.64}{40} > .5$, that is $\overline{x} > .541$.

So $\frac{\text{number preferring A}}{400} > .541$. So,

$$\text{number preferring A} > 216.4$$

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?

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?

answer: A 95% confidence means about $5\% = 1/20$ will be wrong. You'd expect about 2 to be wrong.

With a probability $p = .05$ of being wrong, the number wrong follows a Binomial$(40, p)$ distribution. This has expected value 2, and standard deviation $\sqrt{40(.05)(.95)} = 1.38$. 10 wrong is $(10\text{-}2)/1.38 =$ 5.8 standard deviations from the mean. This would be surprising.

χ^2 confidence interval

A statistician chooses 27 randomly selected dates, and when examining the occupancy records of a particular motel for those dates, finds a standard deviation of 5.86 rooms rented. If the number of rooms rented is normally distributed, find the 95% confidence interval for the standar deviation of the number of rooms rented.

Solution

answer: We have $n = 27$ and $s^2 = 5.86$. If we fix a hypothesis for σ^2 we know

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

We used R to find the critical values. (Or use the χ^2 table.)

```
c025 = qchisq(.975,26) = 41.923
c975 = qchisq(.025,26) = 13.844
```

The 95% confidence interval for σ^2 is

$$\left[\frac{(n-1)\cdot s^2}{c_{.025}}, \frac{(n-1)\cdot s^2}{c_{.975}}\right] = \left[\frac{26\cdot 5.86}{41.923}, \frac{26\cdot 5.86}{13.844}\right] = [3.6343, 11.0056$$

We can take square roots to find the 95% confidence interval for σ

$$[1.9064, 3.3175]$$

Linear regression (least squares)

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

2. You have trivariate date (x_1, x_2, y): $(1, 2, 3)$, $(2, 3, 5)$, $(3, 0, 1)$. Set up a least squares fit of the multiple regression model
$y = ax_1 + bx_2 + c$.

3. Redo problem (2) for general data $(x_{i,1}, x_{i,2}, y_i)$.

18.05 Introduction to Probability and Statistics
Spring 2014