Biomedical Informatics (BMI) 713

Computational Statistics for Biomedical Sciences

Fall, 2017; Tues & Thur, 10-11:30, Modell 100A

Lecture 1: Probability Distributions

BMI 713
October 18, 2017
Peter J Park

Acknowledgments:
Kim Gauvreau
Vince Carey

Teaching Assistants:
Eric Bartell, ebartell@g.harvard.edu
Jacob Luber, jluber@g.harvard.edu
Giorgio Melloni, Giorgio_Melloni@hms.harvard.edu
Tiziana Sanavia, Tiziana_Sanavia@hms.harvard.edu

Lab: There will be weekly laboratory sessions (times TBD) for programming exercises and reviewing lecture material. There will also be drop-by hours each week to get help with debugging one's code.

Prerequisites: No previous knowledge in statistics or programming is required, although those with no programming experience will be expected to devote a significant amount of extra time. If you are not familiar with R, extra help will be available.

Grading: Weekly assignments: 70%; Final: 30%. The course may be taken Pass/Fail if your program allows it.

Auditing: If there is space, anyone, including postdoctoral fellows, may audit the course with the permission of the course director.

Lab hours?
Extra help for R programming?

Course Outline:

Week	Content
Week 1 (10/19)	• Probability distributions • Expectation and variance
Week 2 (10/24, 10/26)	• Sampling distribution • Confidence intervals
Week 3 (10/31, 11/2)	• Two-sample tests • Non-parametric tests
Week 4 (11/7, 11/9)	• Contingency tables • Correlation analysis
Week 5 (11/14, 11/16)	• Linear regression • Multiple linear regression; survival analysis
Week 6 (11/21)	• P-values revisited • Multiple testing; false discovery rate
Week 7 (11/28, 11/30)	• Bayesian methods • Paper discussions
Week 8 (12/5, 12/7)	• Paper discussions • Final exam

Your goals for this class?

Statistical Inference

- Methods used for drawing conclusions about a population based on the information contained in a sample of observations
- We need to introduce some basic principles of probability to establish foundation for statistical inference

- We investigate the properties of a sample mean
- We extrapolate findings from sample data to the larger population using the methods of confidence intervals and hypothesis testing
- We extend to the comparison of two (or more) means

Mean and variance

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{(x_1 + x_2 + x_3 + \cdots + x_n)}{n}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Why *(n-1)* instead of *n* in the denominator?

Random Variables

- Any quantity or characteristic that can assume a number of different values such as that any particular outcome is determined by chance
- It can be **discrete** (countable number of outcomes) or **continuous** (any value in a specified interval)
- Represent a potential outcome of the random variable X by x
- Probability mass function:

$$0 < P(X = x) \leq 1$$

$$\sum P(X = x) = 1$$

Expected value

- If a random variable is able to take on a large number of values, a probability mass function is not the most useful way to summarize its behavior
- Instead, we calculate measures of location and dispersion
- The average value assumed by a random variable is called its **expected value** (or **population mean**)
- Represented by $E(X)$ or μ
- Supposed a random variable X is able to take on the k distinct values, $x_1, x_2, \ldots, x_k$

$$E(x) = \sum_{i=1}^{k} x_i P(X = x_i)$$

Variance

- The variance of a random variable X is called the population variance and is represented by Var(x) or σ^2
- It quantifies the dispersion of the possible outcomes of X around the expected value μ

$$Var(X) = E\left[(X-\mu)^2\right]$$

$$= \sum_{i=1}^{k} (x_i - \mu)^2 P(X = x_i)$$

$$= \sum_{i=1}^{k} x_i^2 P(X = x_i) - \mu^2$$

Example:

- Let X be a r.v. that represents the number of diagnostic services that a child receives during an office visit.

x	$\mathsf{P}(X = x)$
0	0.671
1	0.229
2	0.053
3	0.031
4	0.010
5	0.006

$$\begin{aligned} \mathrm{E}(X) &= 0(0.671) + 1(0.229) + 2(0.053) \\ &\quad + 3(0.031) + 4(0.010) + 5(0.006) \\ &= 0.498 \end{aligned}$$

$$\begin{aligned} \sigma^2 &= [0^2(0.671) + 1^2(0.229) + 2^2(0.053) \\ &\quad + 3^2(0.031) + 4^2(0.010) \\ &\quad + 5^2(0.006)] - (0.498)^2 \\ &= 0.782 \end{aligned}$$

$$\begin{aligned} \sigma &= \sqrt{0.782} \\ &= 0.884 \end{aligned}$$

The binomial distribution

- Consider the dichotomous random variable Y, taking on one of two possible values, e.g., "failure" and "success"
- This type of r.v. is called **Bernoulli random variable**
- Example: Let Y represent the disease status of a person exposed to the hepatitis B virus
- $Y = 1$, if the person develops hepatitis; $Y=0$, if he/she does not

- Suppose five people were infected with the virus. Let X be the r.v. that represents the number of persons who develop disease
- The probability distribution of X is called a **binomial distribution**

Back to the example

- Suppose that 30% of individuals who are exposed to the virus develop disease (p=0.3). Y_i = 1, if the i^{th} person develops disease; Y_i = 0, otherwise.
- Consider n=2. Let X be the r.v. that represents the numbers of persons who develop disease.
- What is the probability that X=0, 1, or 2?

The binomial distribution

- Given *n* independent outcomes of a Bernoulli random variable *Y*, each with probability of success *p.*
- X is the total number of successes
- The probability of exactly *k* successes is

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Aside: permutation and combinations

- Ex) In how many ways can A, B, and C be ordered?

- 3 choices for the first position, 2 choices for the second,and 1 choice for the last position: 3 x 2 x 1 = 6
- n! ("n factorial") = n x (n-1) x (n-2) x … x 2 x 1
- Ex) In how many ways can 3 letters be selected out of the first 6 letters when the order of selection matters?

Combinations

- What if order does not matter?
- 3 letters can be ordered in 3! ways. Therefore, the number of ways in which 3 letters can be selected out of 6 when the order of selection does not matter is 120/6 = 20

$$
\begin{aligned}
{}_n\mathrm{C}_k &= \binom{n}{k} \\
&= \frac{{}_n\mathrm{P}_k}{k!} \\
&= \frac{n(n-1)\times\cdots\times(n-k+1)}{k(k-1)\times\cdots\times(2)(1)} \\
&= \frac{n!}{k!\,(n-k)!}
\end{aligned}
$$

Back to the hepatitis example

- Suppose 5 are exposed. Recall p=0.3. What is the probability that 2 of them will develop disease?

- In the same example, what is the probability that ***at most 2*** of them will develop disease?

Mean and variance of a binomial r.v.

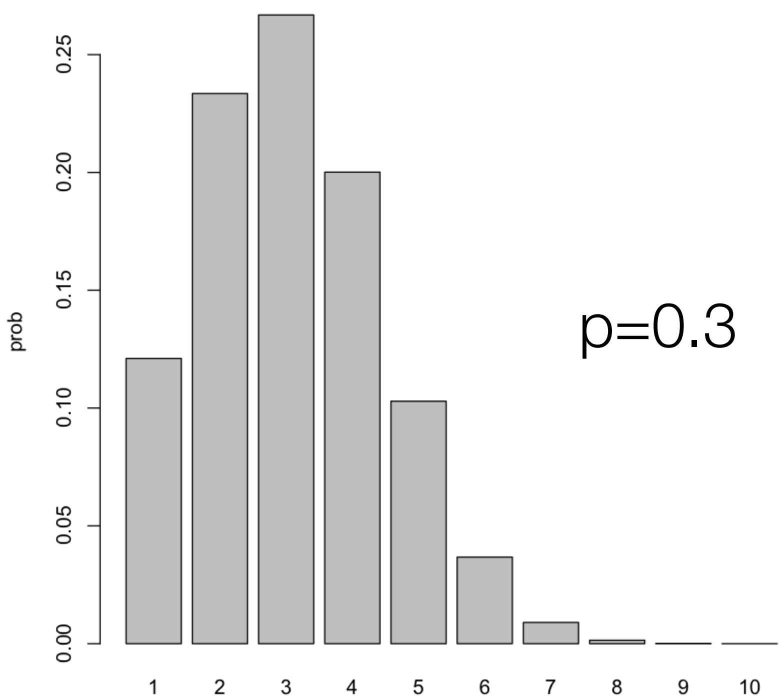
- A binomial distribution can be summarized in terms of a measure of location and measure of dispersion.
- Let $q = 1\text{-}p$

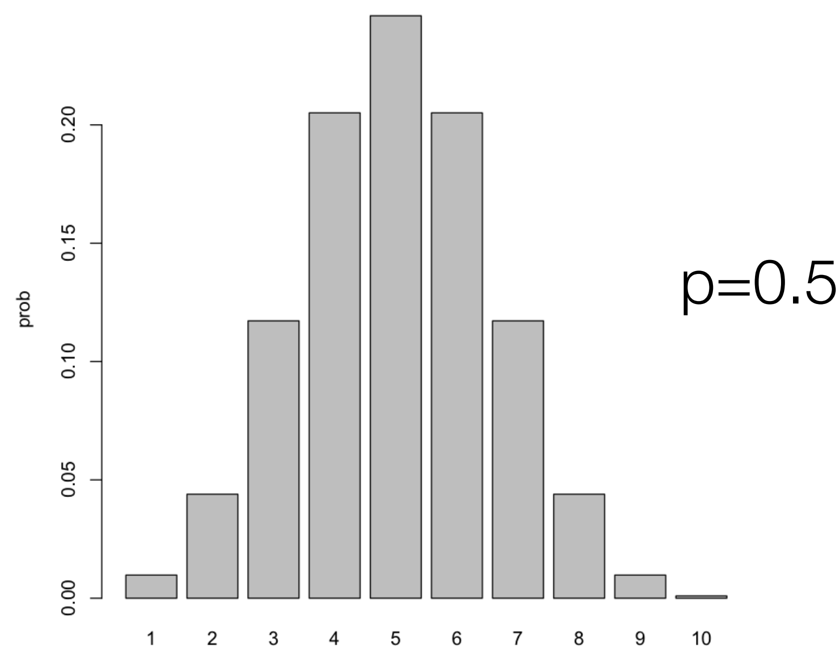
$$
\begin{aligned}
E(x) &= \sum_{i=1}^{m} x_i P(X = x_i) \\
&= \sum_{k=0}^{n} k \binom{n}{k} p^k q^{n-k} \\
&= np
\end{aligned}
$$

$$
\begin{aligned}
Var(X) &= E\left[(X - \mu)^2\right] \\
&= \sum_{i=1}^{m} (x_i - \mu)^2 P(X = x_i) \\
&= \sum_{k=0}^{n} (k - np)^2 \binom{n}{k} p^k q^{n-k} \\
&= npq
\end{aligned}
$$

Back to the example

- Suppose repeated samples of size 10 are selected from those exposed to hepatitis B.
- The mean number of disease per sample: $np = (10)(0.3) = 3$
- The variance: $npq = (10)(0.3)(0.7) = 2.1$. s.d. = sqrt(2.1) = 1.45





The Binomial Distribution

Description

Density, distribution function, quantile function and random generation for the binomial distribution with parameters `size` and `prob`.

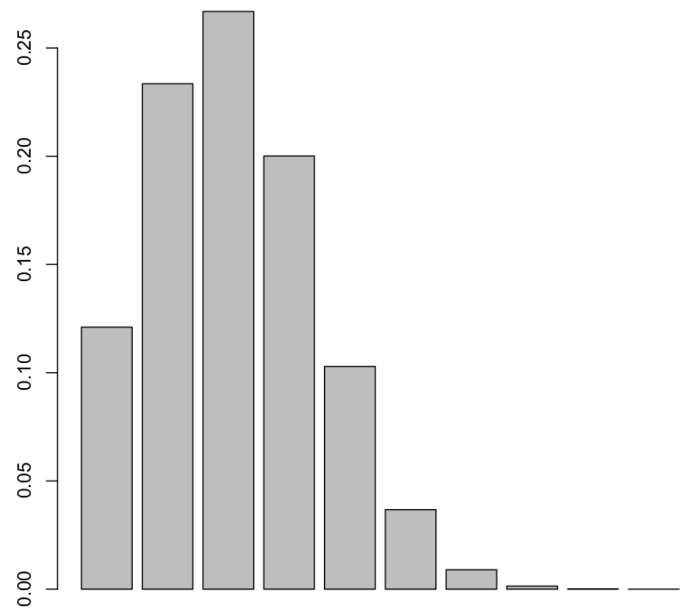
This is conventionally interpreted as the number of 'successes' in `size` trials.

Usage

```
dbinom(x, size, prob, log = FALSE)
pbinom(q, size, prob, lower.tail = TRUE, log.p = FALSE)
qbinom(p, size, prob, lower.tail = TRUE, log.p = FALSE)
rbinom(n, size, prob)
```

"density"	"probability distribution"	"quantile function"
$P(X = x_i)$	$P(X \leq x_i)$	$Q(p) = \{x \mid P(X \leq x) = p\}$

- In a previous example, n=2, p=0.3.
- P(X=0) = (1-p)(1-p) = (0.7)2 = 0.49
- `dbinom(0,2,prob=.3)`
- P(X=1) = p(1-p)+(1-p)p = 0.42
- `dbinom(1,2,prob=.3)`



- `barplot(dbinom(1:10,10,prob=.3))`

Continuous probability distributions

- A continuous r.v. can take any value in a specified interval
- The probability distribution of X is represented by a smooth curve called a **probability density function (pdf)**
- The total area under the pdf is 1.

$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$

- The prob. associated with any one specific value is 0: P(X=a) = 0
- The **cumulative distribution function (cdf)** of X is

$$\begin{aligned} F(a) &= P(X \leq a) \\ &= \int_{-\infty}^{a} f(x)\, dx \end{aligned}$$

Expectation and variance

- The expected value E(X) is the average value taken on by the random variable X:

$$E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx$$

- Variance is the average squared distance of each possible value of X from μ

$$\begin{aligned} Var(X) &= E\left[(X-\mu)^2\right] \\ &= \int_{-\infty}^{\infty} (x-\mu)^2\, f(x)\, dx \end{aligned}$$

- Standard deviation of X: $\sigma = \sqrt{Var(X)}$

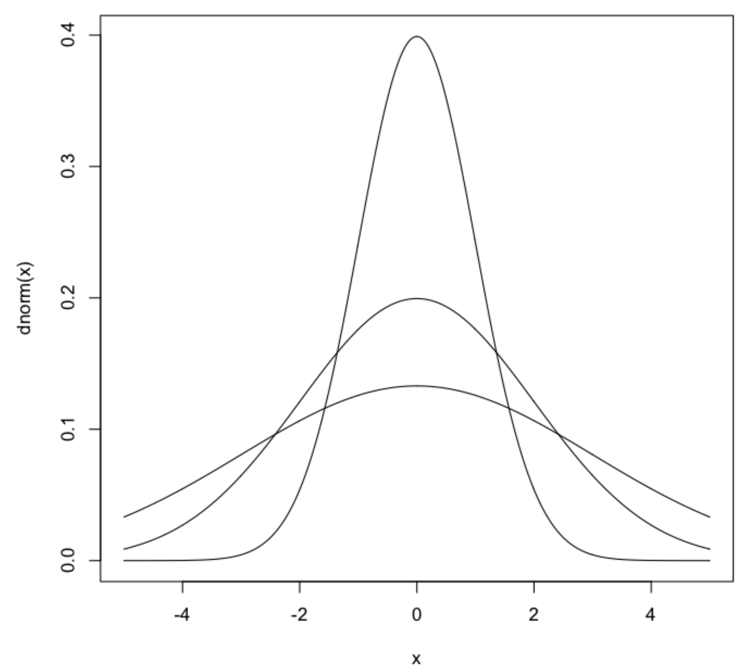
The normal distribution

- Also called “Gaussian distribution”
- The probability density function of a normal random variable X is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad -\infty < x < \infty$$

$\mu = E(X)$ and $\sigma^2 = Var(X)$

```
x=seq(-5,5,.01)
plot(x,dnorm(x),type="n")
lines(x,dnorm(x))
lines(x,dnorm(x,sd=2))
lines(x,dnorm(x,sd=3))
```



- The normal distribution with mean and variance is represented by $N(\mu, \sigma^2)$

- To find $P(X \leq b)$, we would have to draw the probability density function of $N(\mu, \sigma^2)$ and determine the area to the left of *b*
- The standard normal distribution: N(0,1)

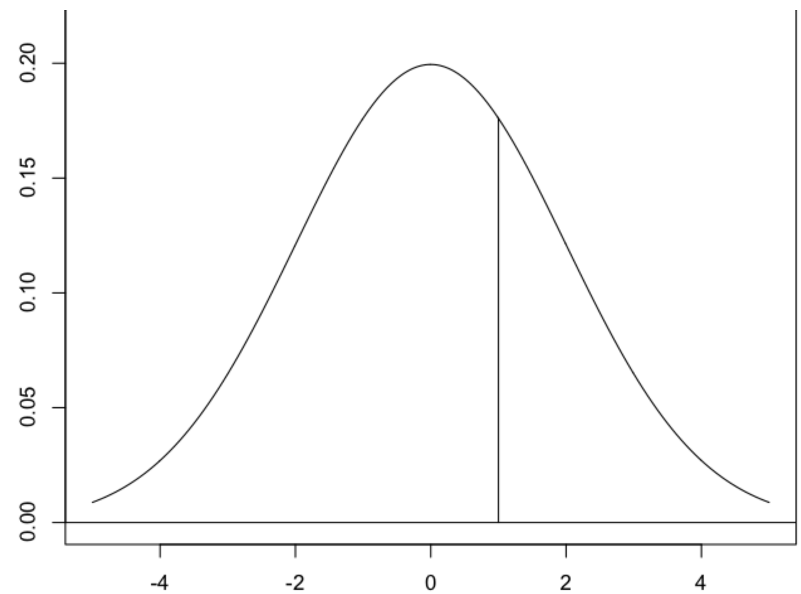
$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

- Useful approximations for the standard normal:
  - (-1,1) contains 68% of the area under the curve
  - (-2, 2) contains 95%, (-2.5, 2.5) contains 99%

- The cumulative distribution function for a standard normal curve is represented by

$$\Phi(x) = P(X \leq x)$$

- X~N(0,1)
  - P(X>2)?
  - P(-2<X<2)?




```
dnorm(x, mean = 0, sd = 1, log = FALSE)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
rnorm(n, mean = 0, sd = 1)
```

$$Q(p) = \{x | P(X \leq x) = p\}$$

The standard normal

If $X \sim N(\mu, \sigma^2)$ and

$$Z = \frac{X - \mu}{\sigma},$$

then $Z \sim N(0, 1)$.

Example

- Suppose the expression levels of gene X is normally distributed with mean = 100 and variance = 225 across my cohort.
- What is the probability that the expression of X in a randomly selected sample is above 120?

$$
\begin{aligned}
P(X > 120) &= P\left(\frac{X - 100}{15} > \frac{120 - 100}{15}\right) \\
&= P(Z > 4/3) \\
&= 0.0912
\end{aligned}
$$

- A **z-score** quantifies how far the value of interest lies from the mean, measured in units of the standard deviation