Review of Probability

Corresponds to Chapter 2 of Tamhane and Dunlop

Slides prepared by Elizabeth Newton (MIT),
with some slides by Jacqueline Telford
(Johns Hopkins University)

Concepts (Review)

A population is a collection of all units of interest.

A sample is a subset of a population that is actually observed.

A measurable property or attribute associated with each unit of a population is called a variable.

A parameter is a numerical characteristic of a population.

A statistic is a numerical characteristic of a sample.

Statistics are used to infer the values of parameters.

A random sample gives a non-zero chance to every unit of the population to enter the sample.

In probability, we assume that the population and its parameters are **known** and compute the probability of drawing a particular sample.

In statistics, we assume that the population and its parameters are **unknown** and the sample is used to infer the values of the parameters.

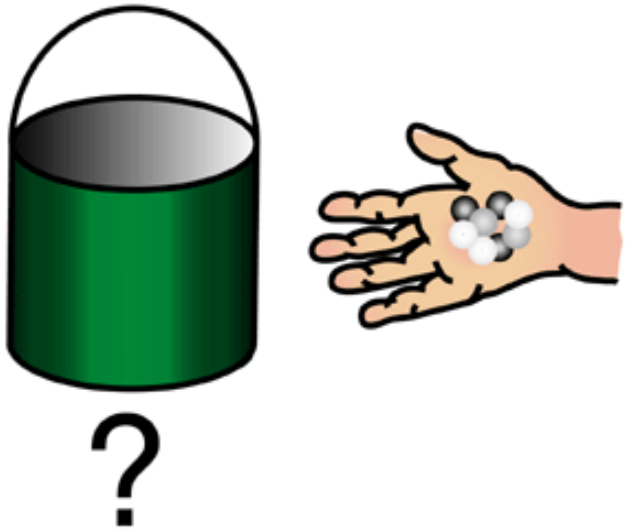
Different samples give different estimates of population parameters (called sampling variability).

Sampling variability leads to “sampling error”.

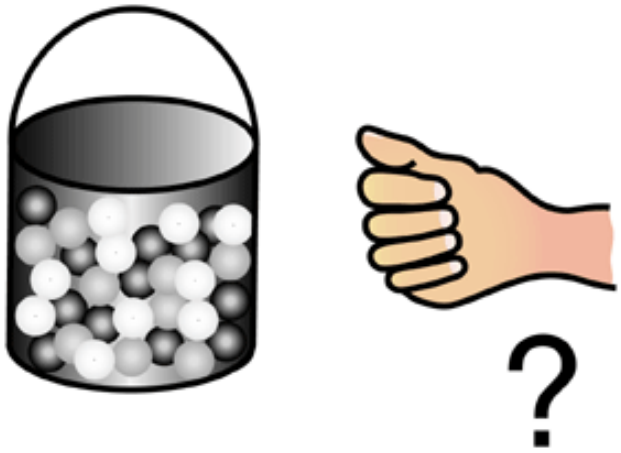
Probability is deductive (general -> particular)

Statistics is inductive (particular -> general)

Difference between Statistics and Probability



Statistics: Given the information in your hand, what is in the box?



Probability: Given the information in the box, what is in your hand?

Based on: *Statistics*, Norma Gilbert, W.B. Saunders Co., 1976.

Probability Concepts

Random experiment – procedure whose outcome cannot be predicted in advance. E.g. toss a coin twice

Sample Space (S) – The finest grain, mutually exclusive, collectively exhaustive listing of all possible outcomes (Drake, *Fundamentals of Applied Probability Theory*)
S={H,H},{H,T},{T,H},{T,T}

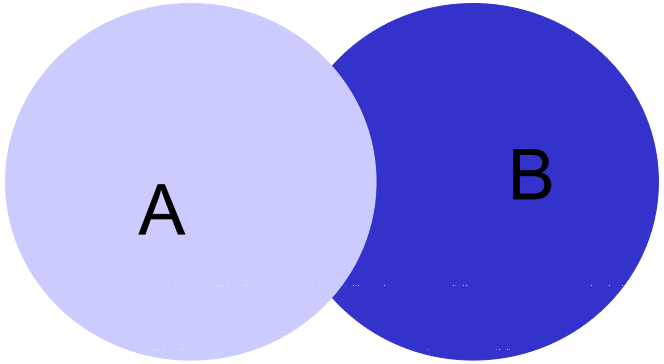
Event (A) a set of outcomes (subset of S). E.g. No heads
A={T,T}

Union (or) E.g. A=heads on first, B=heads on second
A U B= {H,T},{H,H},{T,H}

Intersection (and): E.g. A= heads on first, B=heads on second
$A \cap B$ = {H,H}

Complement of Event A – set of all outcomes not in A. E.g.
A={T,T}, A^c={H,H},{H,T},{T,H}

Venn Diagram



Axioms of Probability

Associated with each event A in S is the probability of A, P(A)

Axioms:

1. **$P(A) \geq 0$**
2. **P(S) = 1 where S is the sample space**
3. **P(A U B) = P(A) + P(B) if A and B are mutually exclusive**

E.g. P(ace or king) = P(ace)+P(king)=1/13+1/13=2/13.

Theorems about probability can be proved using these axioms and these theorems can be used in probability calculations.

$P(A) = 1 - P(A^c)$ (see "birthday problem" on p. 13)

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

E.g. P(ace or black) = P(ace) + P(black) – P(ace and black)

= 4/52 + 26/52 – 2/52 = 28/52 = 7/13

Conditional Probabiity:

$P(A|B) = P(A \cap B)/P(B)$

$P(A \cap B) = P(A|B)P(B)$

E.g. Drawing a card from a deck of 52 cards, P(Heart)=1/4.

However, if it is known that the card is red, P(Heart | Red) = ½.

Sample space has been reduced to the 26 red cards.

(See page 16)

Independence
P(A|B)=P(A)

There are situations in which knowing that event B occurred gives no information about event A, E.g. knowing that a card is black gives no information about whether it is an ace.

P(ace | black) = 2/26 = 4/52 = P(ace).

If two events are independent then $P(A \cap B)=P(A)P(B)$

$P(A \cap B)=P(A|B)P(B)=P(A)P(B)$

E.g. P(ace of hearts) = P(ace) * P(hearts) = 4/52 * 13/52 = 1/52

Independent events are not the same as disjoint events.

Strong dependence between disjoint events.

E.g. card is red means can't be black. P(A|B)=0.

Summary

If A and B are disjoint:

$P(A \cup B) = P(A) + P(B)$

$P(A \cap B) = 0$

If A and B are independent:

$P(A \cap B) = P(A) * P(B)$

$P(A \cup B) = P(A) + P(B) – P(A \cap B)$

Bayes Theorem

- $P(A \cap B) = P(A|B)\,P(B) = P(B|A)\,P(A)$
- $P(B|A) = P(A|B)\,P(B) / P(A)$
- $P(B)$ = prior probability
- $P(B|A)$ = posterior probability

- E.g. P(heart | red)=P(red | heart) * P(heart) / P(red) = 1* 0.25 / 0.5 = 0.5

- Monte Hall problem (page 20)

Sensor Problem

Assume that there are two chemical hazard sensors: A and B.

Let P(A falsely detecting a hazardous chemical)=0.05 and the same for B.

What is the probability of both sensors falsely detecting a hazardous chemical?

$P(A \cap B) = P(A|B) \times P(B) = P(A) \times P(B) = 0.05 \times 0.05 = 0.0025$

- **only** if A and B are independent (use different detection methods).

If A and B are both "fooled" by the same chemical substance, then $P(A \cap B) = P(A \mid B) \times P(B) = 1 \times 0.05 = 0.05$

- which is **20** times the rate of false alarms (same type of sensor)

DON'T assume independence without good reason!

HIV Testing Example

Made-up data

	HIV +	HIV -	
Test positive (+)	95	495	590
Test negative (-)	5	9405	9410
	100	9900	10000

P(HIV +) = 100/10000 = .01 (prevalence)

P(Test + | HIV +) = 95/100 = 0.95 (sensitivity)

P(Test - | HIV -) = 9405/9900 = .95 (specificity)

P(Test - | HIV +) = 5/100 = .05 (false negatives)

P(Test + | HIV -) = 495/9900 = .05 (false positives)

want these to be high

want these to be low

P(HIV + | Test +) =

95/590 = 0.16

This is one reason why we don't have mass HIV screening

Suggestions for Solving Probability Problems

Draw a picture

- Venn diagram
- Tree or event diagram (Probabilistic Risk Assessment)
- Sketch

Write out all possible combinations if feasible

Do a smaller scale problem first

- Figure out the algorithm for the solution
- Increment the size of the problem by one and check algorithm for correctness
- Generalize algorithm (mathematical induction)

Counting rules

Number of Possible Arrangements of Size r from n Objects:

	Without Replacement	With Replacement
Ordered:	$\frac{n!}{(n-r)!}$	n^r
Unordered:	$\binom{n}{r}$	$\binom{n+r-1}{r}$

Source: Casella, George, and Roger L. Berger. *Statistical Inference*. Belmont, CA: Duxbury Press, 1990, page 16.

Counting rules (from Casella & Berger)

For these examples, see pages 15-16 of: Casella, George, and Roger L. Berger. *Statistical Inference*. Belmont, CA: Duxbury Press, 1990.

Birthday Problem

At a gathering of s randomly chosen students what is the probability that at least 2 will have the same birthday?

P(at least 2 have same birthday)=

1-P(all s students have different birthdays).

Assume 365 days in a year. Think of students' birthdays as a sample of these 365 days.

The total number of possible outcomes is:

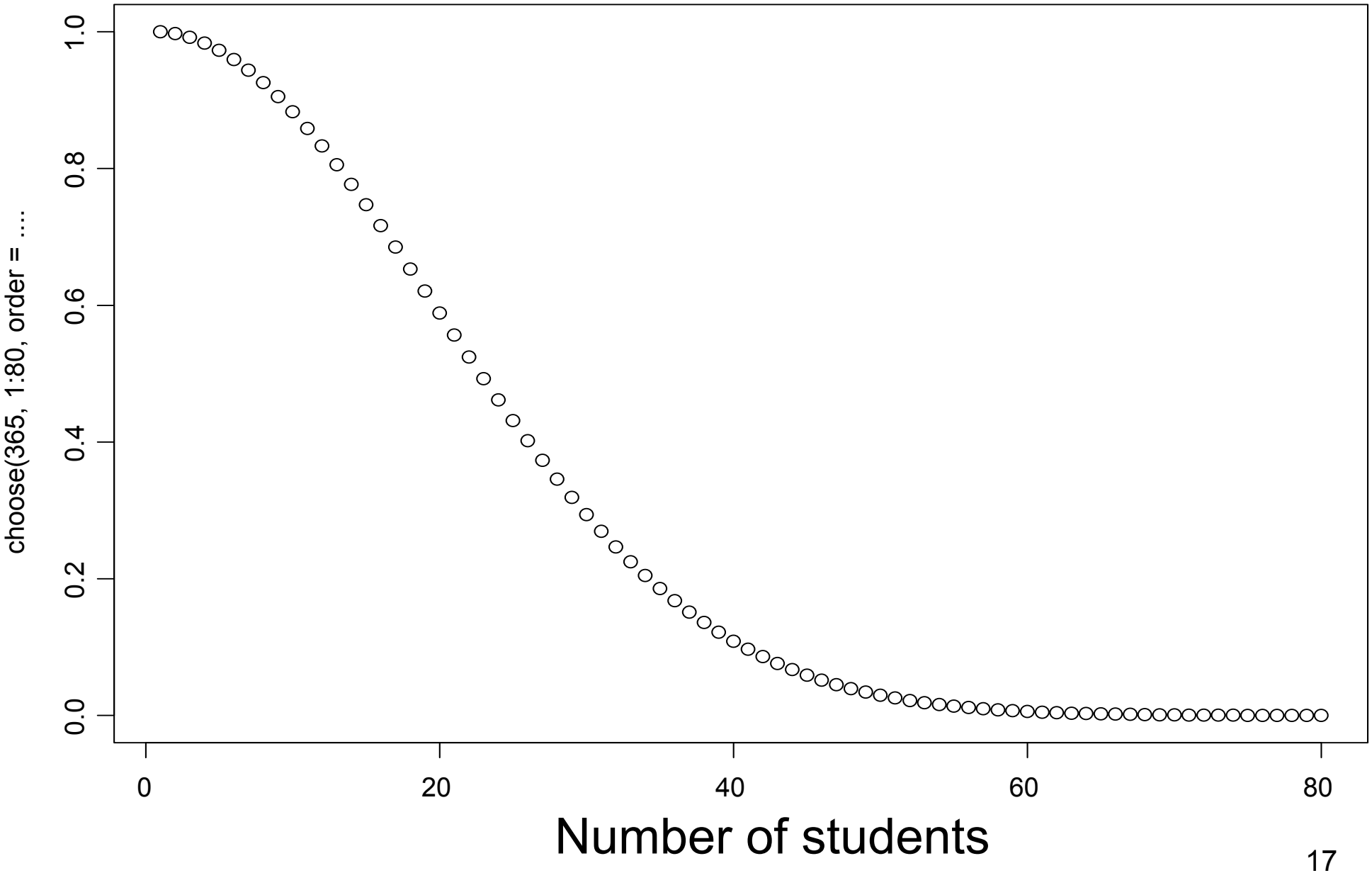
$N=365^s$ (ordered, with replacement)

The number of ways that s students can have different birthdays is

M=364!/(365-s)! (ordered, without replacement)

P(all s students have different birthdays) is M / N.

Probability that all students have different birthdays



See “Harry Potter and the Sorcerer’s Stone” by J.K. Rowling.

Another Counting Rule

The number of ways of classifying n items into k groups with r_i in group i, $r_1+r_2+\ldots+r_k=n$, is:

$n! / (r_1! \, r_2! \, r_3! \ldots r_k!)$

For example: How many ways are there to assign 100 incoming students to the 4 houses at Hogwarts?

(1.6 * 10^57)

Random Variables

A random variable (r.v.) associates a unique numerical value with each outcome in the sample space

Example:

$$X = \begin{cases} 1 & \text{if coin toss results in a head} \\ 0 & \text{if coin toss results in a tail} \end{cases}$$

Discrete random variables: number of possible values is finite or countably infinite: $x_1, x_2, x_3, x_4, x_5, x_6, \ldots$

Probability mass function (p.m.f.)

$f(x) = P(X = x)$ (Sum over all possible values =1 always)

Cumulative distribution function (c.d.f)

$$F(x) = P(X \leq x) = \sum_{k \leq x} f(k)$$

- See Table 2.1 on p. 21 (p.m.f. and c.d.f. for sum of two dice)
- See Figure 2.5 on p. 22 (p.m.f. and c.d.f. graphs for two dice)

Continuous Random Variables

An r.v. is continuous if it can assume any value from one or more intervals of real numbers

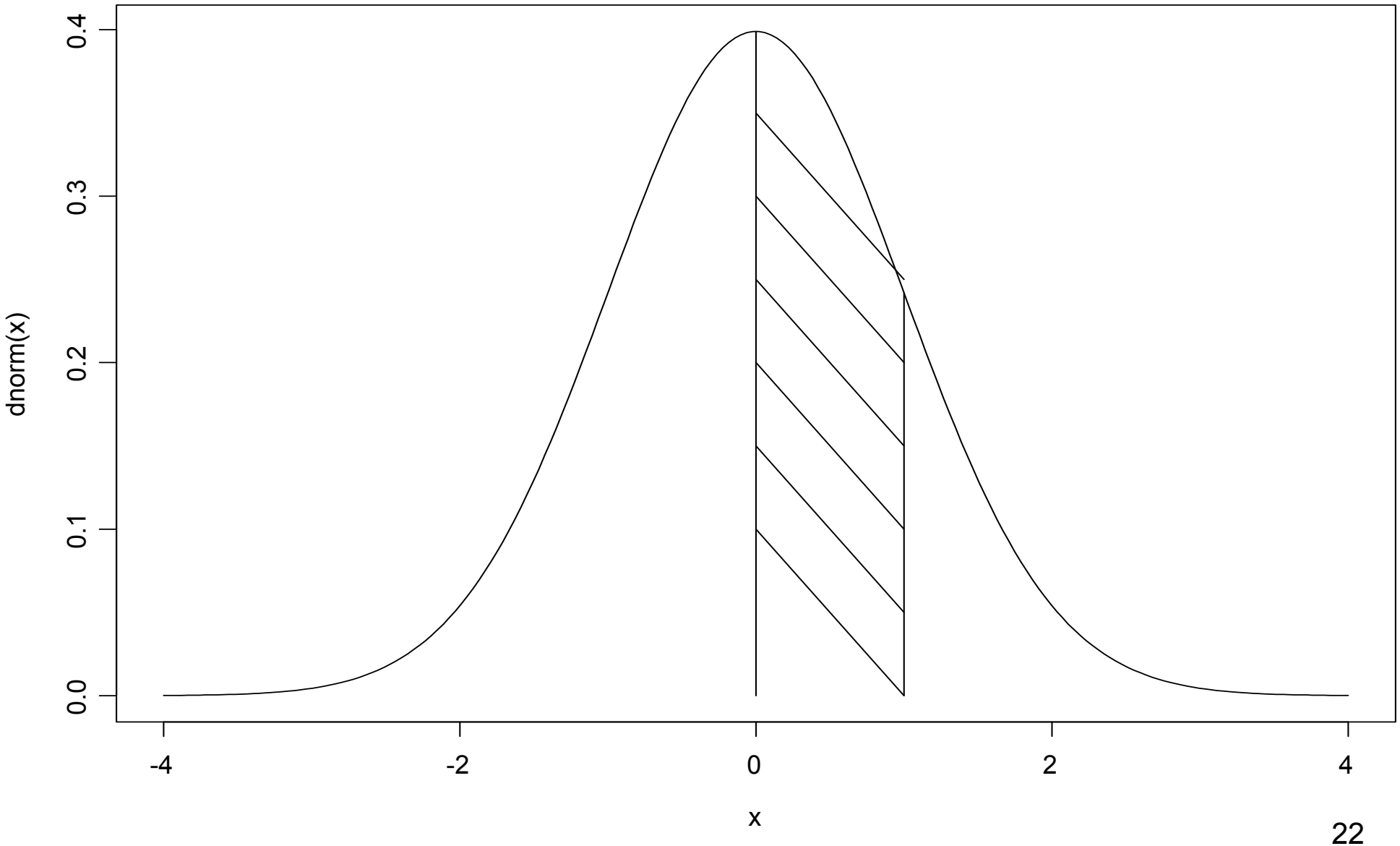
Probability density function (p.d.f.) f(x)

$$f(x) \geq 0$$

$$\int_{-\infty}^{\infty} f(x)dx = 1 \quad \text{(Area under the curve} = 1 \text{ always)}$$

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)ds \quad \text{for any } a \leq b$$

P(0<X<1) for standard normal= area under curve between 0 and 1



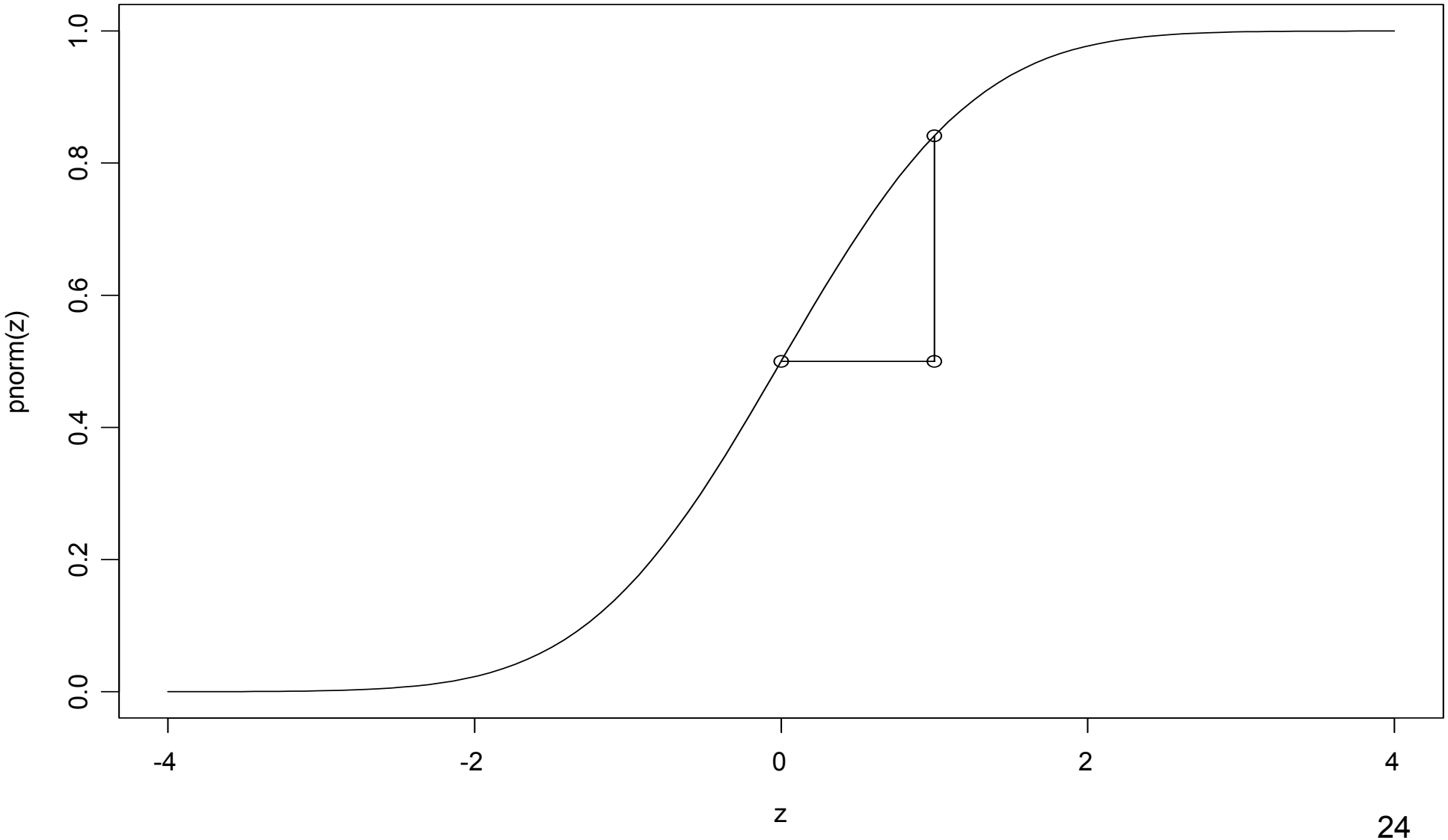
Cumulative Distribution Function

The cumulative distribution function (c.d.f.), denoted *F(x)*, for a continuous random variable is given by:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$$

$$f(x) = \frac{dF(x)}{dx}$$

P(0<Z<1) for standard normal= F(1)-F(0) =0.8413-0.5 = 0.3413 (table page 674)



Expected Value

The expected value or mean of a discrete r.v. X,denoted by E(X), μ_x, or simply μ, is defined as:

$$E(X) = \mu = \sum_x x\ f(x) = x_1 f(x_1) + x_2 f(x_2) + \ldots$$

This is essentially a weighted average of the possible values the r.v. can assume, weights=f(x)

The expected value of a continuous r.v. X is defined as:

$$E(X) = \mu = \int x\ f(x) dx$$

Variance and Standard Deviation

The variance of an r.v. X, denoted by Var(X), σ_x^2, or simply σ^2, is defined as:

$$\text{Var}(X) = \sigma^2 = E[(X - \mu)^2]$$

$$\text{Var}(X) = E[(X - \mu)^2] = E(X^2 - 2\mu X + \mu^2)$$

$$= E(X^2) - 2\mu E(X) + E(\mu^2)$$

$$= E(X^2) - 2\mu\mu + \mu^2$$

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

The standard deviation (SD) is the square root of the variance. Note that the variance is in the square of the original units, while the SD is in the original units.

- See Example 2.17 on p. 26 (mean and variance of two dice)

Quantiles and Percentiles

For $0 \le p \le 1$ the p^{th} quantile (or the $100p^{th}$ percentile), denoted by θ_p, of a continuous r.v. X is defined by the following equation:

$$P(X \le \theta_p) = F(\theta_p) = p$$

$\theta_{.5}$ is called the median

- See Example 2.20 on p. 30 (exponential distribution)

Jointly distributed random variables and independent random variables

See pp. 30-33

Joint Distributions

For a discrete distribution:

$f(x,y) = P(X=x, Y=y)$

$f(x,y) \geq 0$ for all x and y

$\sum_x \sum_y f(x,y)=1$

Marginal Distributions

- $g(x) = P(X=x) = \sum_y f(x,y)$
- $h(y) = P(Y=y) = \sum_x f(x,y)$

- Independent if joint distribution factors into product of marginal distributions
- $f(x,y) = g(x)\, h(y)$

Conditional Distributions

f(y|x) = f(x,y) / g(x)

If X and Y are independent:

f(y|x) = g(x) h(y) / g(x) = h(y)

Conditional distribution is just a probability distribution defined on a reduced sample space.

For every x, $\sum_y f(y|x) = 1$

Covariance and Correlation

$$\text{Cov}(X,Y) = \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - E(X)E(Y)$$
$$= E(XY) - \mu_X \mu_Y$$

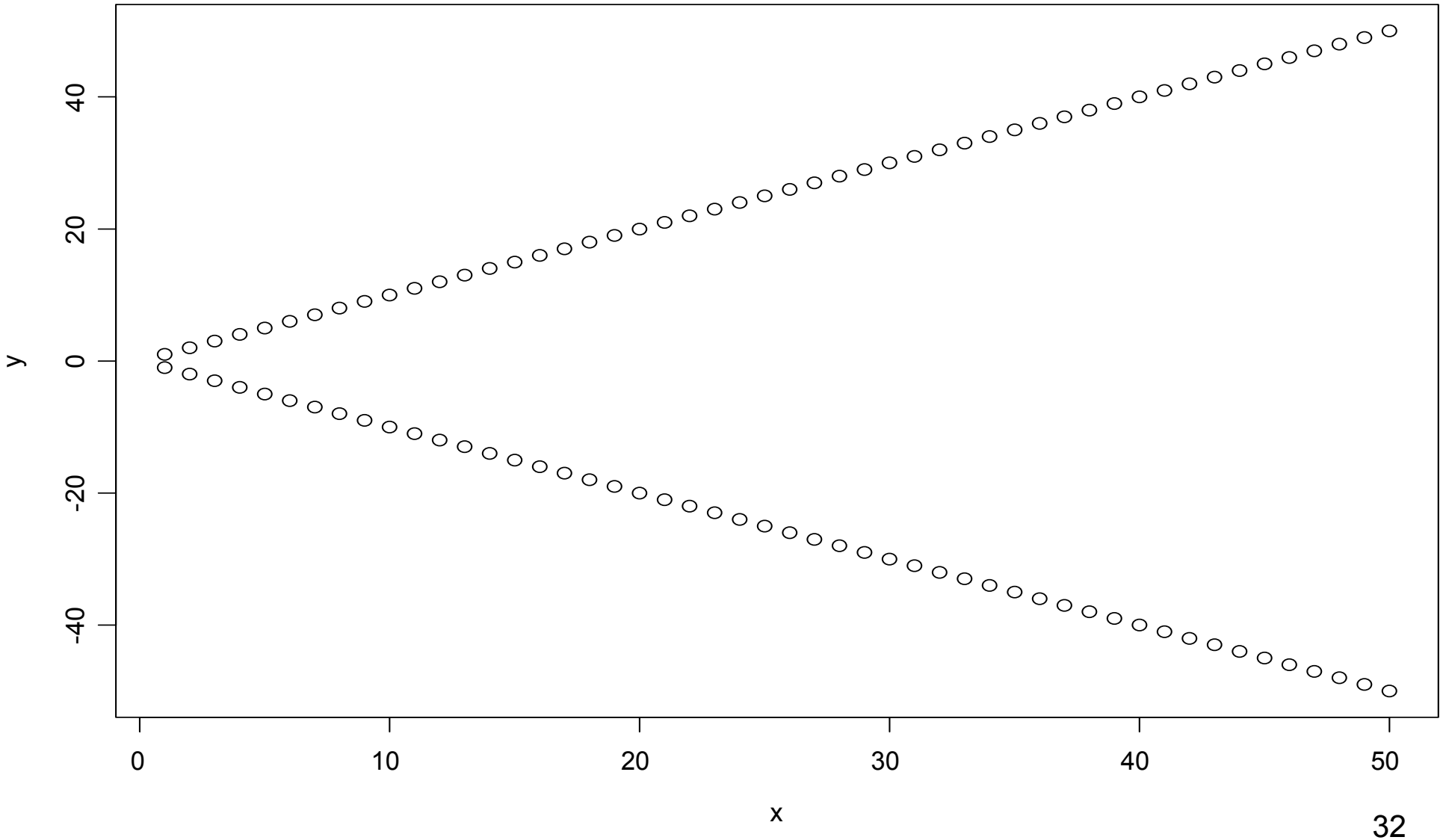
If X and Y are independent, then E(XY) = E(X)E(Y) so the covariance is zero. The other direction is not true.

Note that: $$E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\, y\, f(x,y)\, dx\, dy$$

$$\rho_{XY} = corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\sigma_{XY}}{\sigma_x \sigma_Y}$$

- See Examples 2.26 and 2.27 on pp. 37-38 (prob vs. stat grades)

Example 2.25 in text
y=x with probability 0.5 and y= -x with probability 0.5
y is not independent of x, yet covariance is zero



Two Famous Theorems

Chebyshev's Inequality: Let $c > 0$ be a constant. Then, irrespective of the distribution of X,

$$P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}$$

- See Example 2.29 on p. 41 (exact vs. Cheb. for two dice)

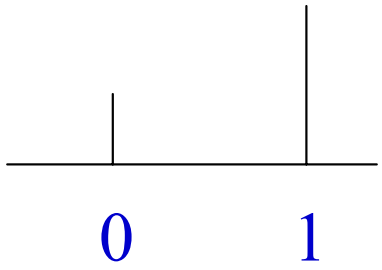
Weak Law of Large Numbers: Let $\overline{X}$ be the sample mean of n i.i.d. observations from a population with finite mean μ and variance σ^2. Then, for any fixed $c > 0$,

$$P(|\overline{X} - \mu| \geq c) \to 0 \text{ as } n \to \infty$$

Selected Discrete Distributions

Bernoulli trials: (single coin flip)

$$f(x) = P(X = x) = \begin{cases} p & \text{if x} = 1 \text{ (success)} \\ 1 - p & \text{if x} = 0 \text{ (failure)} \end{cases}$$



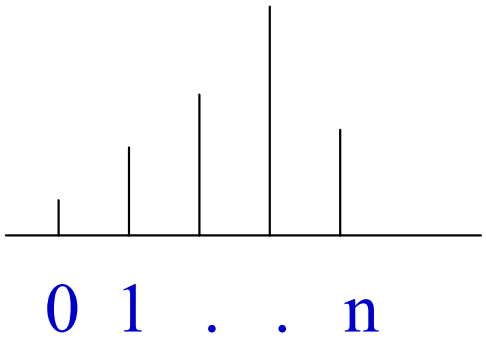
E(X) = *p* and Var(X) = *p(1-p)*

Binomial distribution: (multiple coin flips)

X successes out of *n* trials

$$f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, \ldots, n$$

E(X) = *np* and Var(X) = *np(1-p)*



- See Example 2.30 on p. 43 (teeth)

Selected Discrete Distributions (cont)

Hypergeometric: drawing balls from the box without replacing the balls (as in the hand with the question mark)

Poisson: number of occurrences of a rare event

Geometric: number of failures before the first success

Multinomial: more than two outcomes

Negative Binomial: number of trials to get r successes

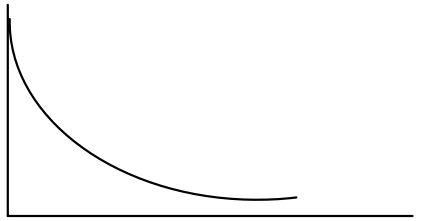
Uniform: N equally likely events

- See Table 2.5, p. 59 for properties of these distributions

Selected Continuous Distributions

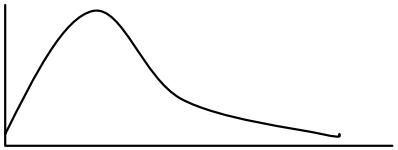
Uniform: equally likely over an interval

Exponential: lifetimes of devices with no wear-out ("memoryless"), interarrival times when the arrivals are at random

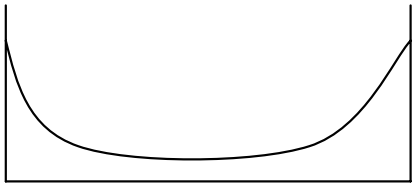


Gamma: used to model lifetimes, related to many other distributions

Lognormal: lifetimes (similar shape to Gamma but with longer tail)



Beta: not equally likely over an interval



- See Table 2.5, p. 59 for properties of these distributions

Normal Distribution

First discovered by **de Moivre** (1667-1754) in 1733

Rediscovered by **Laplace** (1749-1827) and also by

Gauss (1777-1855) in their studies of errors in astronomical measurements.

Often referred to as the **Gaussian** distribution.

Carl Friedrick Gauss (1777 - 1855)

Photograph courtesy of John L. Telford, John Telford Photography. Used with permission. Currency from 1991.

Karl Pearson (1857 - 1936)

"Many years ago I called the Laplace-Gauss curve the *NORMAL* curve, which name, while it avoids an international question of priority, has the disadvantage of leading people to believe that all other distributions of frequency are in one sense or another *ABNORMAL*.

That belief is, of course, not justifiable."

Karl Pearson, 1920

Normal Distribution ("Bell-curve", Gaussian)

A continuous r.v X has a normal distribution with parameter μ and σ^2 if its probability density function is given by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-(x-\mu)^2 / 2\sigma^2] \quad \text{for } -\infty < x < \infty$$

E(X) = μ and Var(X) = σ^2 (see Figure 2.12, p. 53)

Standard normal distribution: $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$

- See Table A.3 on p. 673 $\Phi(z) = P(Z \le z)$

$$P(X \le x) = P\left(Z = \frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma} = z\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

- See Examples 2.37 and 2.38 on pp. 54-55 (computations)

Percentiles of the Normal Distribution

Suppose that the scores on a standardized test are normally distributed with mean 500 and standard deviation of 100. What is the 75th percentile score of this test?

$$P(X \le x) = P\left(\frac{X-500}{100} \le \frac{x-500}{100}\right) = \Phi\left(\frac{x-500}{100}\right) = 0.75$$

From Table A.3, $\Phi(0.675) = 0.75$

$$\frac{x-500}{100} = 0.675 \Rightarrow x = 500 + (0.675)(100) = 567.5$$

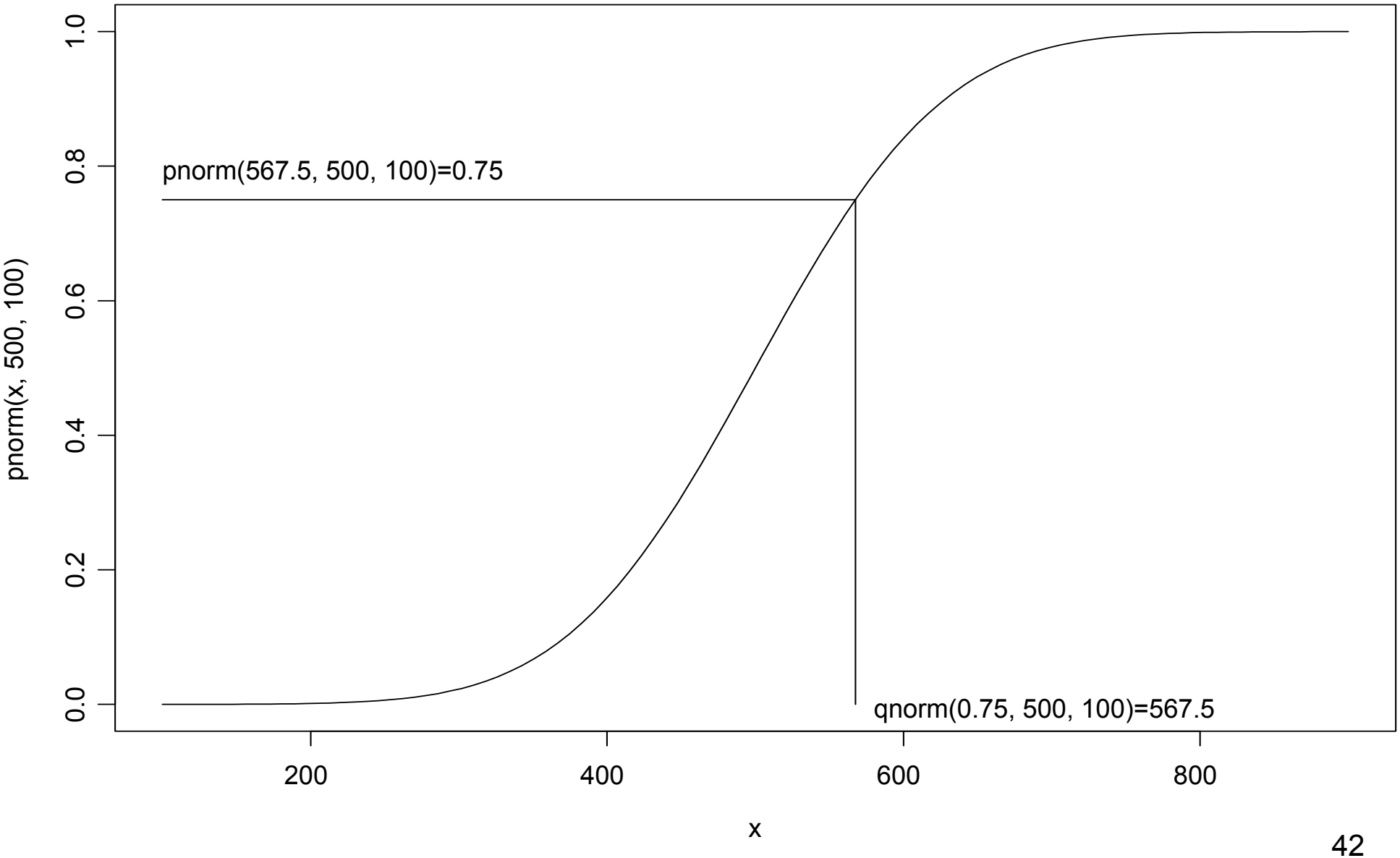
Useful Information about the Normal Distribution:

~68% of a normal population is within $\pm 1\sigma$ of μ

~95% of a normal population is within $\pm 2\sigma$ of μ

~99.7% of a normal population is within $\pm 3\sigma$ of μ

75th percentile for a test with scores which are normally distributed, mean=500, standard deviation=100



Linear Combinations of r.v.s

$X_i \sim N(\mu_i, \sigma_i^2)$ for i = 1, …, n and $Cov(X_i, X_j) = \sigma_{ij}$ for $i \neq j$

Let $X = a_1X_1 + a_2X_2 + \ldots + a_nX_n$ where a_i are constants.

Then X has a normal distribution with mean and variance:

$$E(X) = E(a_1X_1 + a_2X_2 + \ldots + a_nX_n) = a_1\mu_1 + a_2\mu_2 + \ldots + a_n\mu_n = \sum_{i=1}^{n} a_i\mu_i$$

$$Var(X) = Var(a_1X_1 + a_2X_2 + \ldots + a_nX_n) = \sum_{i=1}^{n} a_i^2\sigma_i^2 + 2\sum_{\substack{i=1 \\ i \neq j}}^{n}\sum_{j=1}^{n} a_i a_j \sigma_{ij}$$

$\bar{X}$ = $(X_1 + X_2 + \ldots + X_n) / n$, so $a_i = 1/n$

Therefore, $\bar{X}$ from *n* i.i.d. $N(\mu, \sigma^2)$ observations ~ $N(\mu, \sigma^2/n)$, since the covariances (σ_{ij}) are zero (by independence).