Quantitative Analysis of Clinical Data

Ingo Ruczinski

Associate Professor, Department of Biostatistics, Johns Hopkins University

Office: E3618 SPH   Email: ingo@jhu.edu

http://www.biostat.jhsph.edu/∼iruczins

Logistics

Lectures: M 5:30pm-8:30pm, W2030 SPH

Office hours: By appointment.

Textbooks: [required]

Dawson and Trapp (2002): Basic and clinical biostatistics.
McGraw-Hill 4th edition.

[recommended]

Gonick & Smith (1993): The cartoon guide to statistics.
Collins Reference 1st edition.

Webpage: www.biostat.jhsph.edu/ iruczins/teaching/390.672/

Course learning objectives

⟶ Read, understand, and critically discuss quantitative methods used in the scientific literature on clinical investigation.

⟶ Analyze and interpret basic quantitative data.

Topics covered:

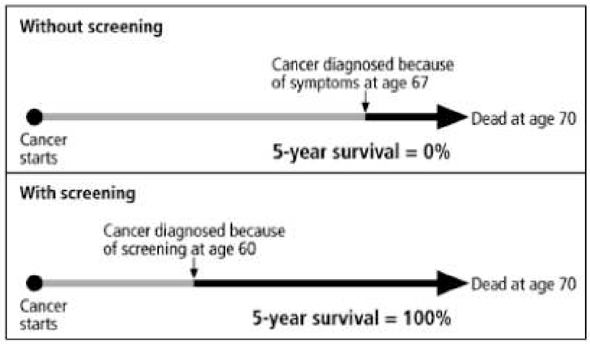
Basic statistical display of data, probabilities and distributions, confidence intervals, tests of hypotheses, likelihood and statistical evidence, tests for goodness of fit, contingency tables, analysis of variance, multiple comparisons, regression and correlation, basic experimental design, observational studies, survival analysis, prediction, methods of evidence-based medicine and decision analysis.

Grading

- This course is **not** offered for credit, it's a *certificate* course.
- The course is pass or fail.
- There are weekly assignments.
- There is a final project (written critique and presentation).

Example 1

Does higher survival rate mean longer life?



$\longrightarrow$ No.

Gigerenzer et. al. (2008)

Example 2

A test with 99% sensitivity and 99% specificity returns a positive result. What is the probability that the person has the disease?

		DISEASE	
		+	−
TEST	+	TP	FP
	−	FN	TN

$\longrightarrow$ It depends.

(On the prevalence of the disease: for example, it is $<$ 10% if the prevalence is 0.1%, 50% if the prevalence is 1%, and $>$ 90% if the prevalence is 10%.)

Example 3

A diagnostic test returns a positive result. A physician might conclude that:

1. The subject probably has the disease.
2. The test result is evidence that the subject has the disease.
3. The subject should be treated for the disease.

Isn't that all the same?

⟶ Not even close.

Example 4

January/February · 2010 IMS Bulletin · 5

Members' Discoveries: *Fatal flaws in cancer research*

Jeffrey Morris, Associate Professor at The University of Texas MD Anderson Cancer Center, reports on the substantial impact of a paper in the *Annals of Applied Statistics:* A recent article published in *The Annals of Applied Statistics (AOAS)* by two MD Anderson researchers—Keith Baggerly and Kevin Coombes—dissects results from a highly-influential series of medical papers involving genomics-driven personalized cancer therapy, and outlines a series of simple yet fatal flaws that raises serious questions about the veracity of the original results. Having immediate and strong impact, this paper, along with related work, is providing the impetus for new standards of reproducibility in scientific research.

In late 2006, investigators at Duke University led by Anil Potti and Joseph Nevins developed genomic signatures to predict sensitivity/resistance to particular chemotherapeutic agents. They built inclusion of genes of unknown origin, and figure duplication. One common error is the swapping of sensitive/resistant labels in the training data, leading to signatures suggesting a therapy to the patients least likely to benefit from it. There are also persistent irregularities in test sample labeling, with some samples mislabeled and others used multiple times in validation, which leads to inaccurate reports of the methods' performance. One study includes four key signature genes whose origins are unknown; they were not produced by their software, and two of them were not even on the microarrays used in the training data. One figure discussed in a recent publication duplicates a figure from an earlier publication dealing with a completely different treatment. Baggerly and Coombes report that when they followed the Duke researchers' approach with these errors corrected, they obtained results no better than chance. reproduce their results with the information given (*The Cancer Letter*, October 23, 2009). As Baggerly notes in an accompanying letter, there has still not been any documented independent validation of the Duke researchers' results.

The warning in Baggerly and Coombes' paper is that clinical trials being conducted were based on questionable scientific results, and that these trials could be putting patients at risk by exposing them to potentially ineffective treatments. Shortly after publication of the *Annals of Applied Statistics* paper, three Duke University clinical trials based on the questionable results were suspended, a fourth trial conducted at Moffit Cancer Center was terminated (*The Cancer Letter*, October 9 and 23, 2009), and a panel of outside experts has been assembled by Duke University to investigate this research and the original results.

Most of the errors discovered in this

Summarizing and Presenting Data

Summary statistics

Location / Center	• mean (average) • median • mode • geometric mean • harmonic mean
Scale	• standard deviation (SD) • inter-quartile range (IQR) • range
Other	• quantile • quartile • quintile

Summary statistics

$$\text{mean} = \frac{1}{n}\sum_{i=1}^{n} x_i = (x_1 + x_2 + \ldots + x_n)/n$$

$$\text{geometric mean} = \sqrt[n]{\prod_{i=1}^{n} x_i} = \exp\left\{\frac{1}{n}\sum_{i=1}^{n}\log x_i\right\}$$

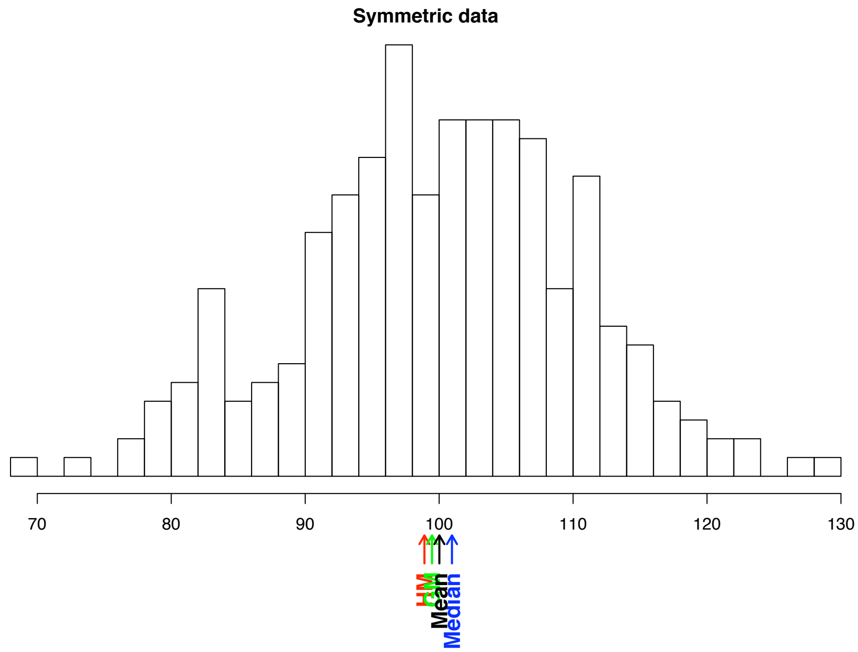
$$\text{harmonic mean} = 1/\left\{\frac{1}{n}\sum_{i=1}^{n}(1/x_i)\right\}$$

$\longrightarrow$ Note: these are all sample means.

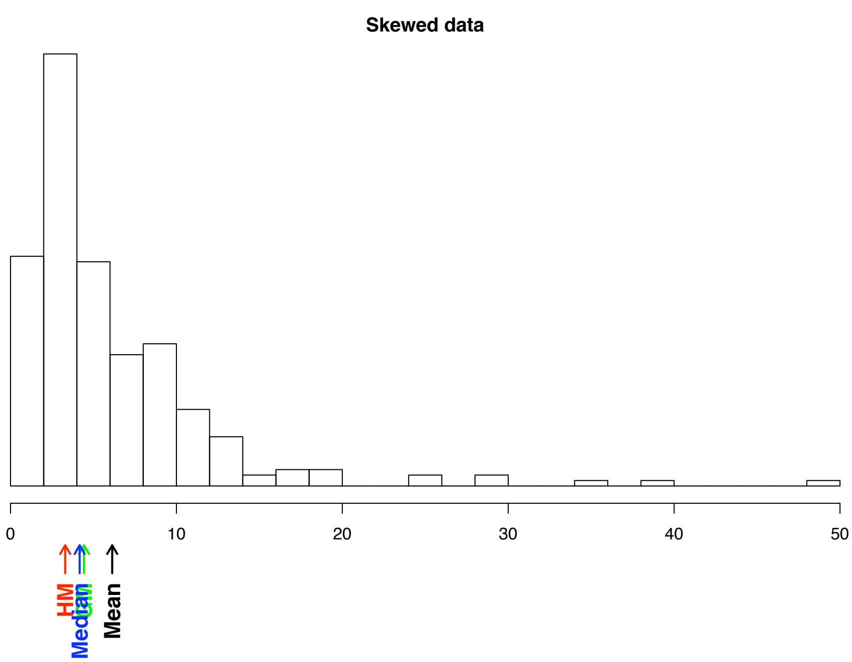
Measures of location / center

- Forget about the mode.
- The mean is sensitive to outliers.
- The median is resistant to outliers.
- The geometric mean is used when a logarithmic transformation is appropriate (for example, when the distribution has a long right tail).
- The harmonic mean may be used when a reciprocal transformation is appropriate (very seldom).

Measures of location / center



Measures of location / center



A key point

The different possible measures of the ”center” of the distribution are all allowable.

You should consider the following though:

$\longrightarrow$ Which is the best measure of the ”typical” value in your particular setting?

$\longrightarrow$ Be sure to make clear which ”average” you use.

Standard deviation (SD)

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

Sample SD $= \sqrt{s^2} = s$

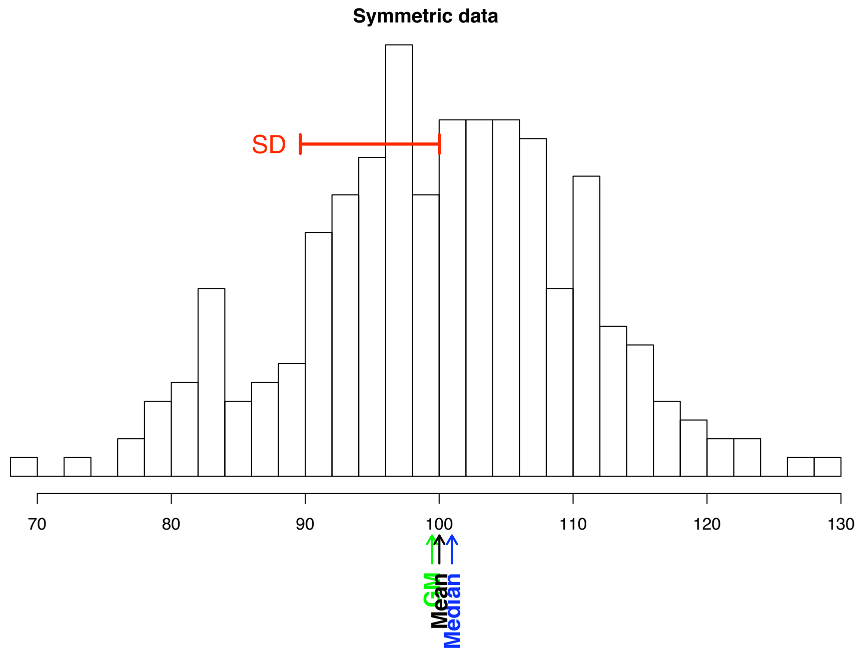
= RMS (distance from average)

= “typical” distance from the average

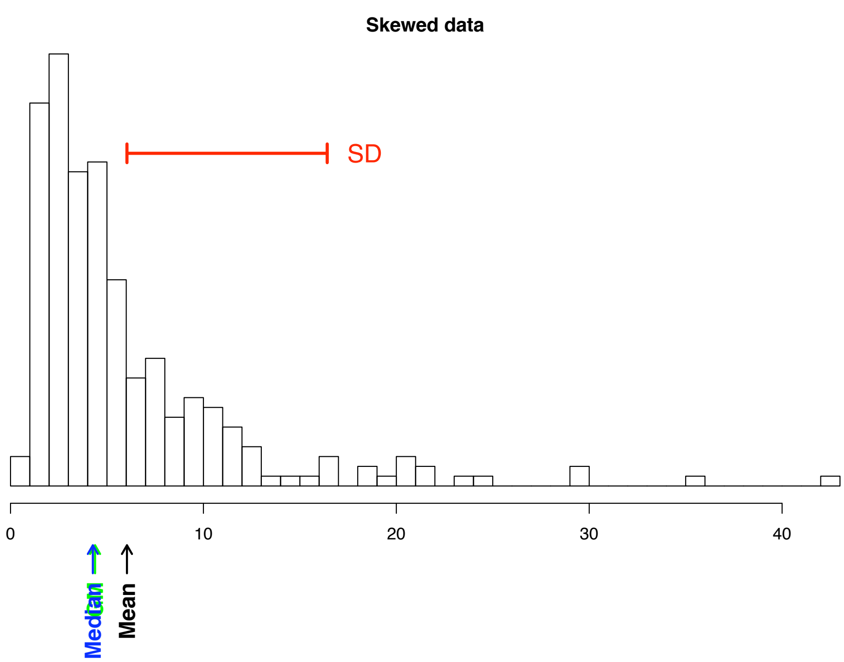
= sort of like ave$\{|x_i - \bar{x}|\}$

$\longrightarrow$ Remember: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

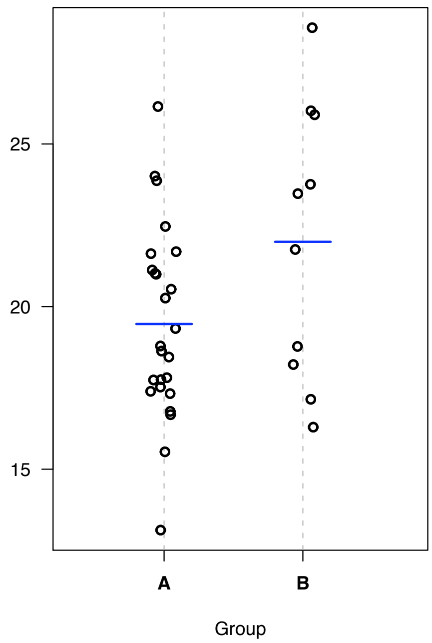
Standard deviation (SD)



Standard deviation (SD)

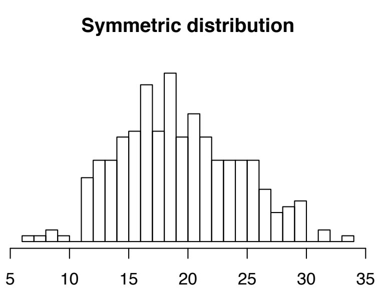


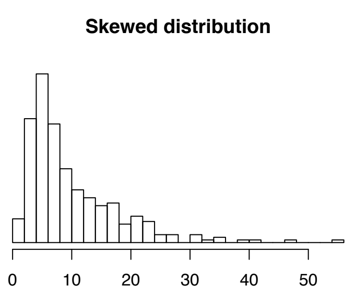
Dotplots



- Few data points per group.
- Possibly many groups.

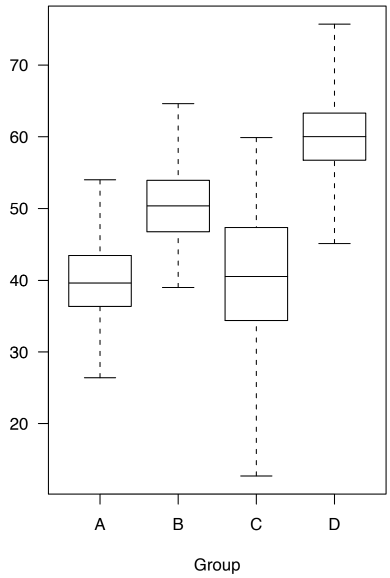
Histograms





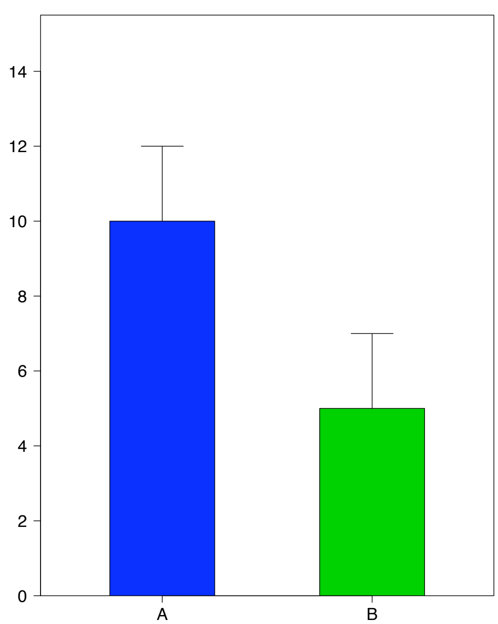
- Many data points per group.
- Few groups.
- Area of the rectangle is proportional to the number of data points in the interval.
- Typically $2\sqrt{n}$ bins is a good choice.

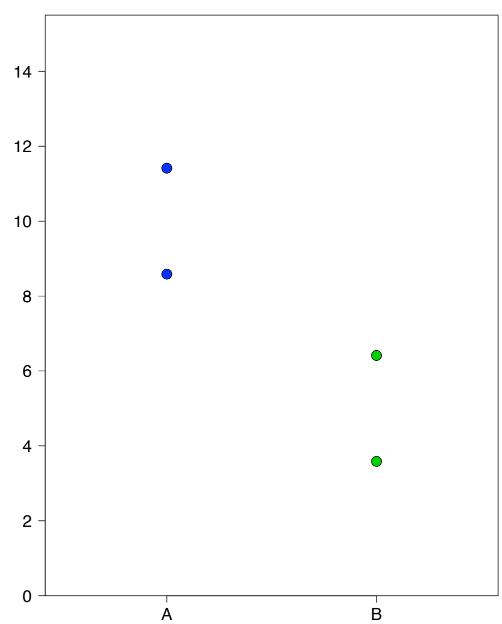
Boxplots



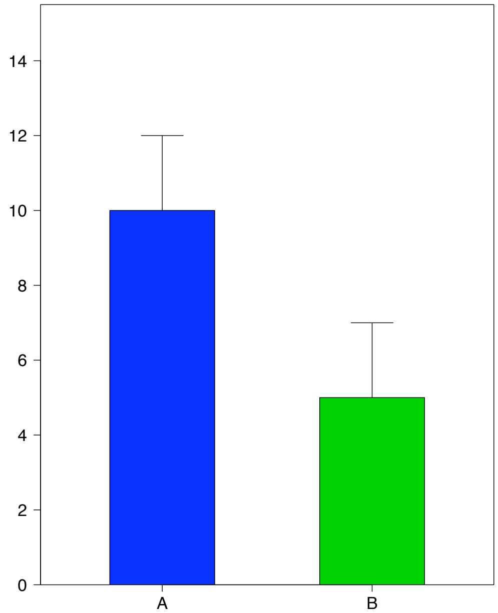
- Many data points.
- Possibly many groups.
- Displays the minimum, lower quartile, median, upper quartile, and the maximum.

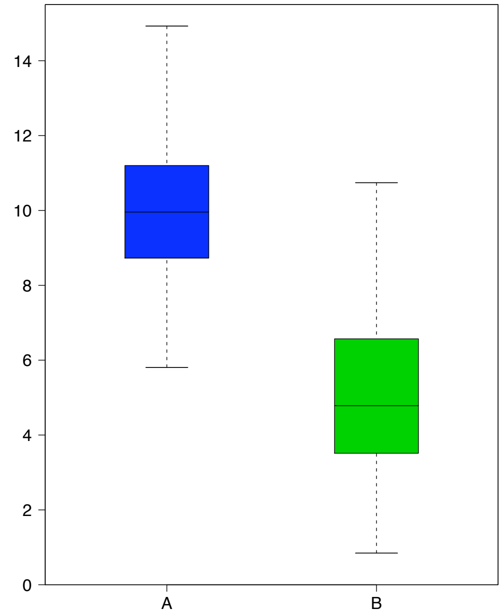
Skyscraper-with-antenna plots



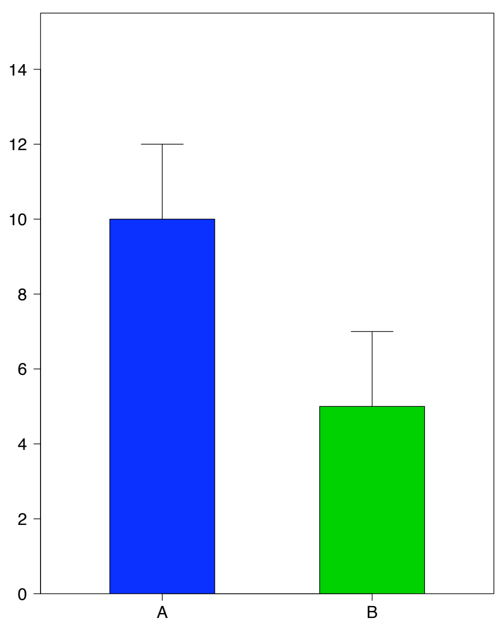


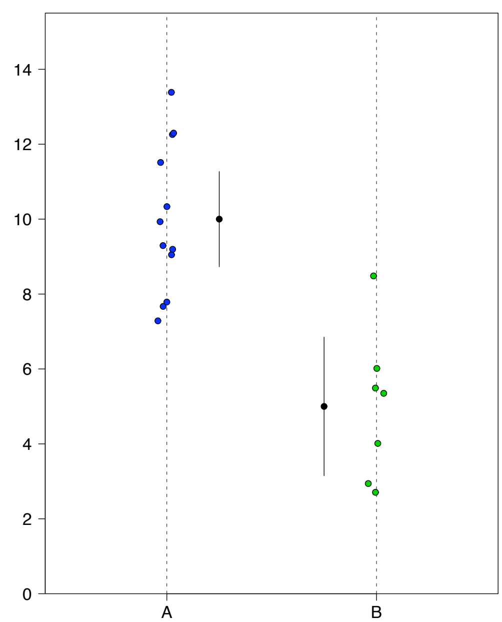
Skyscraper-with-antenna plots



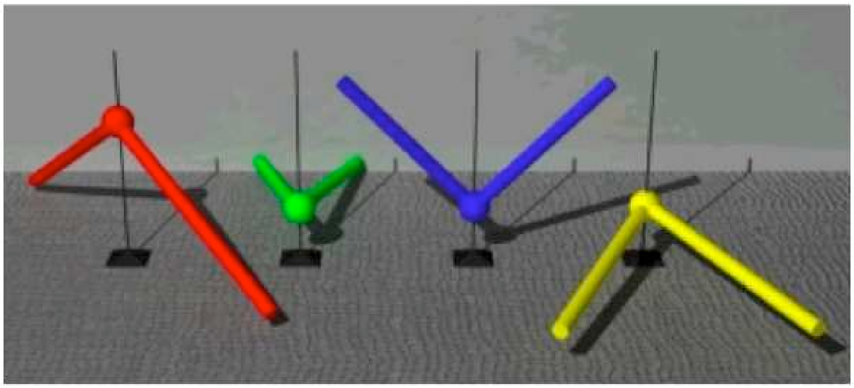


Skyscraper-with-antenna plots

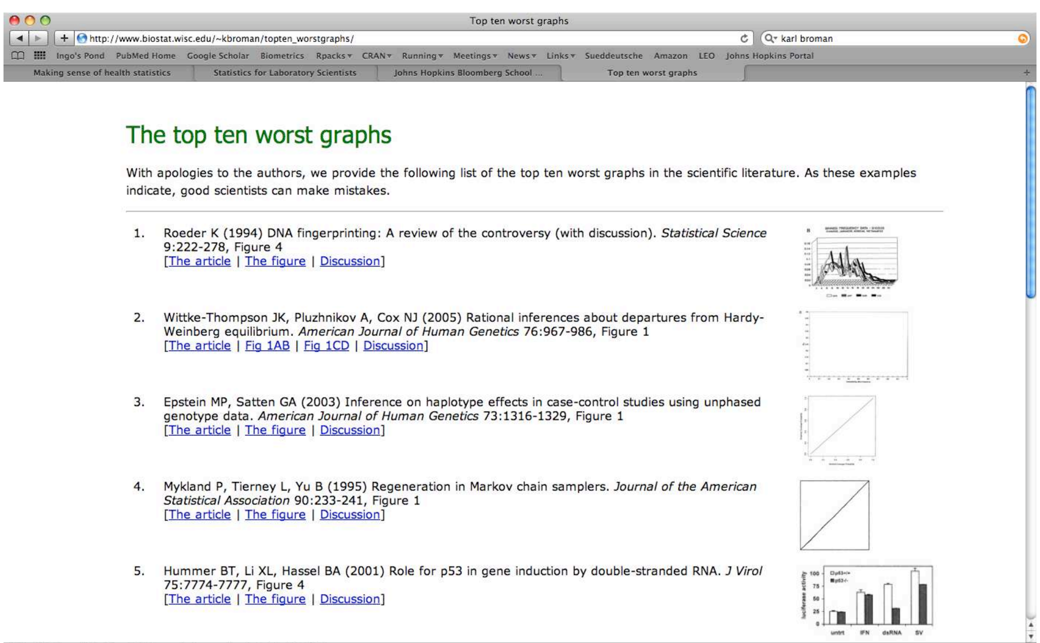




3D graphics



Bad graphs



The top ten worst graphs

With apologies to the authors, we provide the following list of the top ten worst graphs in the scientific literature. As these examples indicate, good scientists can make mistakes.

1. Roeder K (1994) DNA fingerprinting: A review of the controversy (with discussion). *Statistical Science* 9:222-278, Figure 4
[The article | The figure | Discussion]

2. Wittke-Thompson JK, Pluzhnikov A, Cox NJ (2005) Rational inferences about departures from Hardy-Weinberg equilibrium. *American Journal of Human Genetics* 76:967-986, Figure 1
[The article | Fig 1AB | Fig 1CD | Discussion]

3. Epstein MP, Satten GA (2003) Inference on haplotype effects in case-control studies using unphased genotype data. *American Journal of Human Genetics* 73:1316-1329, Figure 1
[The article | The figure | Discussion]

4. Mykland P, Tierney L, Yu B (1995) Regeneration in Markov chain samplers. *Journal of the American Statistical Association* 90:233-241, Figure 1
[The article | The figure | Discussion]

5. Hummer BT, Li XL, Hassel BA (2001) Role for p53 in gene induction by double-stranded RNA. *J Virol* 75:7774-7777, Figure 4
[The article | The figure | Discussion]

Displaying data well

- Let the data speak.

  Show as much information as possible, taking care not to obscure the message.

- Science not sales.

  Avoid unnecessary frills, especially gratuitous colors and 3D.

- In tables, every digit should be meaningful.

  Don't drop ending 0's!

- Be accurate and clear.

Statistics and Probability

What is statistics?

> We may at once admit that any inference from the particular to the general must be attended with some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous, for the nature and degree of the uncertainty may itself be capable of rigorous expression.
>
> — Sir R. A. Fisher

What is statistics?

$\longrightarrow$ Data exploration and analysis.

$\longrightarrow$ Inductive inference with probability.

$\longrightarrow$ Quantification of evidence and uncertainty.

What is probability?

$\longrightarrow$ A branch of mathematics concerning the study of random processes.

Note: Random does not mean haphazard!

What do I mean when I say the following?

The probability that he is a carrier ...

The chance of rain tomorrow ...

$\longrightarrow$ Degree of belief.

$\longrightarrow$ Long term frequency.

The set-up

Experiment

$\rightarrow$ A well-defined process with an uncertain outcome.

Draw 2 balls *with replacement* from an urn containing 4 red and 6 blue balls.

Sample space $\mathcal{S}$

$\rightarrow$ The set of possible outcomes.

{ RR, RB, BR, BB }

Event

$\rightarrow$ A set of outcomes from the sample space (a subset of $\mathcal{S}$).

{the first ball is red} = {RR, RB}

Events are said to occur if one of the outcomes they contain occurs. Probabilities are assigned to events.

Probability rules

$0 \leq \Pr(A) \leq 1$	for any event A
$\Pr(\mathcal{S}) = 1$	where $\mathcal{S}$ is the sample space
Pr(A or B) = Pr(A) + Pr(B)	if A and B are *mutually exclusive*
Pr(not A) = 1 - Pr(A)	complement rule

Example

Study with 10 subjects:

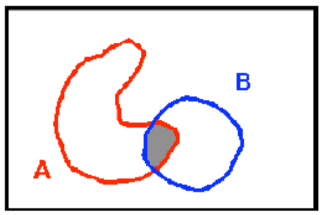
- 2 infected with virus X (only)
- 1 infected with virus Y (only)
- 5 infected with both X and Y
- 2 infected with neither

Experiment: Randomly select one subject (each equally likely).

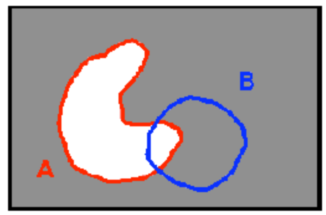
Events:	A = {subject is infected with X}	Pr(A) = 7/10
	B = {subject is infected with Y}	Pr(B) = 6/10
	C = {subject is infected with only X}	Pr(C) = 2/10

Sets

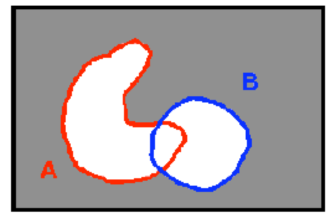
A and B



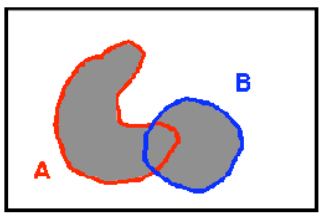
not A



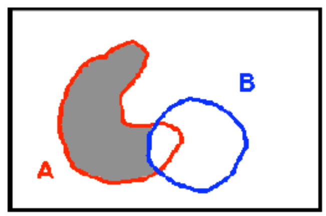
(not A) and (not B)



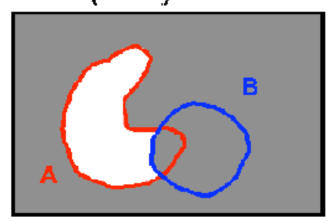
A or B



A and (not B)



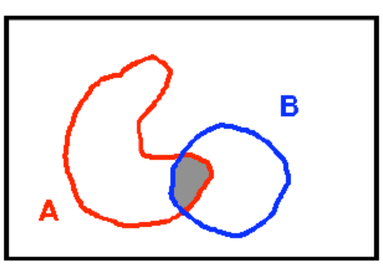
(not A) or B



Conditional probability

Pr(A | B) = *Probability of A given B* = Pr(A and B) / Pr(B)

Example: [2 w/ X only; 1 w/ Y only; 5 w/ both; 2 w/ neither]



A = {infected with X}

B = {infected with Y}

Pr(A | B) = (5/10) / (6/10) = 5/6

Pr(B | A) = (5/10) / (7/10) = 5/7

More rules and a definition

Multiplication rule:

$\longrightarrow$ Pr(A and B) = Pr(A) $\times$ Pr(B $|$ A)

A and B are independent if   Pr(A and B) = Pr(A) $\times$ Pr(B)

If A and B are independent:

$\longrightarrow$ Pr(A $|$ B) = Pr(A)

$\longrightarrow$ Pr(B $|$ A) = Pr(B)

Diagnostics

		DISEASE	
		+	−
TEST	+	TP	FP
	−	FN	TN

Diagnostics

		DISEASE	
		+	−
TEST	+	TP	FP
	−	FN	TN

Sensitivity → Pr (positive test | disease)
Specificity → Pr (negative test | no disease)
Positive Predictive Value → Pr (disease | positive test)
Negative Predictive Value → Pr (no disease | negative test)
Accuracy → Pr (correct outcome)

Diagnostics

		DISEASE	
		+	−
TEST	+	TP	FP
	−	FN	TN

Sensitivity → TP / (TP+FN)
Specificity → TN / (FP+TN)
Positive Predictive Value → TP / (TP+FP)
Negative Predictive Value → TN / (FN+TN)
Accuracy → (TP+TN) / (TP+FP+FN+TN)

Diagnostics

Assume that some disease has a 0.1% prevalence in the population. Assume we have a test kit for that disease that works with 99% sensitivity and 99% specificity. What is the probability of a person having the disease given the test result is positive, if we randomly select a subject from

$\longrightarrow$ the general population?

$\longrightarrow$ a high risk sub-population with 10% disease prevalence?

Diagnostics

		DISEASE	
		+	−
TEST	+	99	999
	−	1	98901

Diagnostics

		DISEASE +	DISEASE −
TEST	+	99	999
	−	1	98901

Sensitivity $\longrightarrow$ 99 / (99+1) = 99%

Specificity $\longrightarrow$ 98901 / (999+98901) = 99%

Positive Predictive Value $\longrightarrow$ 99 / (99+999) $\approx$ 9%

Negative Predictive Value $\longrightarrow$ 98901 / (1+98901) $>$ 99.9%

Accuracy $\longrightarrow$ (99+98901) / 100000 = 99%

Diagnostics

		DISEASE +	DISEASE −
TEST	+	9900	900
	−	100	89100

Diagnostics

DISEASE

TEST	+	−
+	9900	900
−	100	89100

Sensitivity $\longrightarrow$ 9900 / (9900+100) = 99%

Specificity $\longrightarrow$ 89100 / (900+89100) = 99%

Positive Predictive Value $\longrightarrow$ 9900 / (9900+900) $\approx$ 92%

Negative Predictive Value $\longrightarrow$ 89100 / (100+89100) $\approx$ 99.9%

Accuracy $\longrightarrow$ (9900+89100) / 100000 = 99%

Bayes rule

$\longrightarrow$ Pr(A and B) = Pr(A) $\times$ Pr(B | A) = Pr(B) $\times$ Pr(A | B)

$\longrightarrow$ Pr(A) = Pr(A and B) + Pr(A and not B)

= Pr(B) $\times$ Pr(A | B) + Pr(not B) $\times$ Pr(A | not B)

$\longrightarrow$ Pr(B) = Pr(B and A) + Pr(B and not A)

= Pr(A) $\times$ Pr(B | A) + Pr(not A) $\times$ Pr(B | not A)

$\longrightarrow$ Pr(A | B) = Pr(A and B) / Pr(B)

= Pr(A) $\times$ Pr(B | A) / Pr(B)

Bayes rule

Pr(A | B) =

Pr(A) $\times$ Pr(B | A) / Pr(B) =

Pr(A) $\times$ Pr(B | A) / { Pr(A) $\times$ Pr(B | A) + Pr(not A) $\times$ Pr(B | not A) }

Let A denote disease, and B a positive test result!

- $\longrightarrow$ Pr(A | B) is the probability of disease given a positive test result.
- $\longrightarrow$ Pr(A) is the prevalence of the disease.
- $\longrightarrow$ Pr(not A) is 1 minus the prevalence of the disease.
- $\longrightarrow$ Pr(B | A) is the sensitivity of the test.
- $\longrightarrow$ Pr(not B | not A) is the specificity of the test.
- $\longrightarrow$ Pr(B | not A) is 1 minus the specificity of the test.

Random Variables and Distributions

Random variables

Random variable: A number assigned to each outcome of a random experiment.

Example 1: I toss a brick at my neighbor's house.

D = distance the brick travels
X = 1 if I break a window; 0 otherwise
Y = cost of repair
T = time until the police arrive
N = number of people injured

Example 2: Apply a treatment to 10 subjects.

X = number of people that respond
P = proportion of people that respond

Further examples

Example 3: Pick a random student in the School.

S = 1 if female; 0 otherwise
H = his/her height
W = his/her weight
Z = 1 if Canadian citizen; 0 otherwise
T = number of teeth he/she has

Example 4: Sample 20 students from the School

H_i = height of student i
$\overline{H}$ = mean of the 20 student heights
S_H = sample SD of heights
T_i = number of teeth of student i
$\overline{T}$ = average number of teeth

Random variables are . . .

Discrete: Take values in a countable set (e.g., the positive integers).

Example: the number of teeth, number of gall stones, number of birds, number of cells responding to a particular antigen, number of heads in 20 tosses of a coin.

Continuous: Take values in an interval (e.g., [0,1] or the real line).

Example: height, weight, mass, some measure of gene expression, blood pressure.

Random variables may also be partly discrete and partly continuous (for example, mass of gall stones, concentration of infecting bacteria).

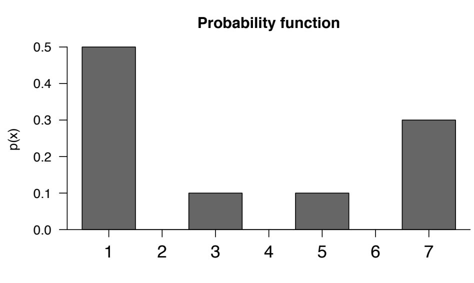
Probability function

Consider a *discrete* random variable, X.

The probability function (or probability distribution, or probability mass function) of X is

$$p(x) = \Pr(X = x)$$

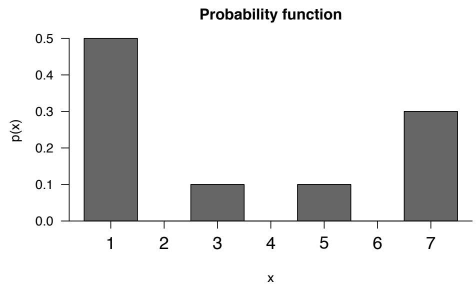
Note that $p(x) \geq 0$ for all x and $\sum p(x) = 1$.



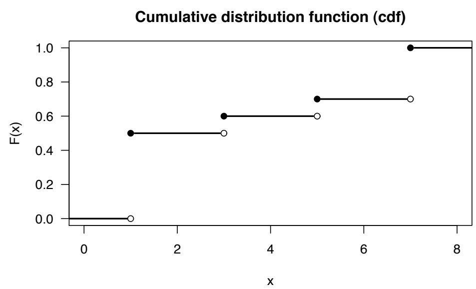
x	p(x)
1	0.5
3	0.1
5	0.1
7	0.3

Cumulative distribution function (cdf)

The cdf of X is $F(x) = \Pr(X \leq x)$



x	p(x)
1	0.5
3	0.1
5	0.1
7	0.3



x	F(x)
$(-\infty,1)$	0
$[1,3)$	0.5
$[3,5)$	0.6
$[5,7)$	0.7
$[7,\infty)$	1.0

Binomial random variable

Prototype: The number of heads in n independent tosses of a coin, where Pr(heads) = p for each toss.
$\rightarrow$ n and p are called *parameters*.

Alternatively, imagine an urn containing red balls and black balls, and suppose that p is the proportion of red balls. Consider the number of red balls in n random draws *with replacement* from the urn.

Example 1: Sample n people at random from a large population, and consider the number of people with some property (e.g., that are graduate students or that have exactly 32 teeth).

Example 2: Apply a treatment to n subjects and count the number of responders (or non-responders).

Example 3: Apply a treatment to 30 groups of 10 subjects. Count the number of groups with at least two responders.

Binomial distribution

Consider the Binomial(n,p) distribution.

That is, the number of red balls in n draws with replacement from an urn for which the proportion of red balls is p.

$\longrightarrow$ What is its probability function?

Example: Let $X \sim$ Binomial(n=9,p=0.2).

$\longrightarrow$ We seek p(x) = Pr(X= x) for x = 0, 1, 2, . . . , 9.

p(0) = Pr(X= 0) = Pr(no red balls) = $(1 - p)^n = 0.8^9 \approx 13\%$.

p(9) = Pr(X= 9) = Pr(all red balls) = $p^n = 0.2^9 \approx 5 \times 10^{-7}$

p(1) = Pr(X= 1) = Pr(exactly one red ball) = . . . ?

Binomial distribution

p(1) = Pr(X= 1) = Pr(exactly one red ball)

= Pr(RBBBBBBBB or BRBBBBBBB or . . . or BBBBBBBBR)

= Pr(RBBBBBBBB) + Pr(BRBBBBBBB) + Pr(BBRBBBBBB)
+ Pr(BBBRBBBBB) + Pr(BBBBRBBBB)
+ Pr(BBBBBRBBB) + Pr(BBBBBBRBB)
+ Pr(BBBBBBBRB) + Pr(BBBBBBBBR)

$= p(1 - p)^8 + p(1 - p)^8 + \ldots p(1 - p)^8 = 9p(1 - p)^8 \approx 30\%$.

How about p(2) = Pr(X= 2)?

How many outcomes have 2 red balls among the 9 balls drawn?

$\longrightarrow$ This is a problem of combinatorics. That is, counting!

Getting at Pr(X= 2)

RRBBBBBBB	RBRBBBBBB	RBBRBBBBB	RBBBRBBBB
RBBBBRBBB	RBBBBBRBB	RBBBBBBRB	RBBBBBBBR
BRRBBBBBB	BRBRBBBBB	BRBBRBBBB	BRBBBRBBB
BRBBBBRBB	BRBBBBBRB	BRBBBBBBR	BBRRBBBBB
BBRBRBBBB	BBRBBRBBB	BBRBBBRBB	BBRBBBBRB
BBRBBBBBR	BBBRRBBBB	BBBRBRBBB	BBBRBBRBB
BBBRBBBRB	BBBRBBBBR	BBBBRRBBB	BBBBRBRBB
BBBBRBBRB	BBBBRBBBR	BBBBBRRBB	BBBBBRBRB
BBBBBRBBR	BBBBBBRRB	BBBBBBRBR	BBBBBBBRR

How many are there?

$9 \times 8 / 2 = 36.$

The binomial coefficient

The number of possible samples of size k selected from a population of size n :

$$\binom{n}{k} = \frac{n!}{k! \times (n-k)!}$$

$\longrightarrow$ $n! = n \times (n-1) \times (n-2) \times \ldots \times 3 \times 2 \times 1$

$\longrightarrow$ $0! = 1$

For a Binomial(n,p) random variable:

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

Example

Suppose Pr(subject responds to treatment) = 90%, and we apply the treatment to 10 random subjects.

$$\begin{aligned}
\text{Pr(exactly 7 subjects respond)} &= \binom{10}{7} \times (0.9)^7 \times (0.1)^3 \\
&= \frac{10 \times 9 \times 8}{3 \times 2} \times (0.9)^7 \times (0.1)^3 \\
&= 120 \times (0.9)^7 \times (0.1)^3 \\
&\approx 5\%
\end{aligned}$$

$$\begin{aligned}
\text{Pr(fewer than 9 respond)} &= 1 - p(9) - p(10) \\
&= 1 - 10 \times (0.9)^9 \times (0.1) - (0.9)^{10} \\
&\approx 26\%
\end{aligned}$$

The world is entropy driven

Assume we are flipping a fair coin (independently) ten times. Let X be the random variable that describes the number of heads H in the experiment.

Pr(TTTTTTTTTT) = Pr(HTTHHHTHTH) = $(1/2)^{10}$

- $\longrightarrow$ There is only one possible outcome with zero heads.
- $\longrightarrow$ There are 210 possibilities for outcomes with six heads.

Thus,

- $\longrightarrow$ $\Pr(X = 0) = (1/2)^{10} \approx 0.1\%$.
- $\longrightarrow$ $\Pr(X = 6) = 210 \times (1/2)^{10} \approx 20.5\%$.

The world is entropy driven

Assume that in a lottery, six out of the numbers 1 through 49 are randomly selected as the winning numbers.

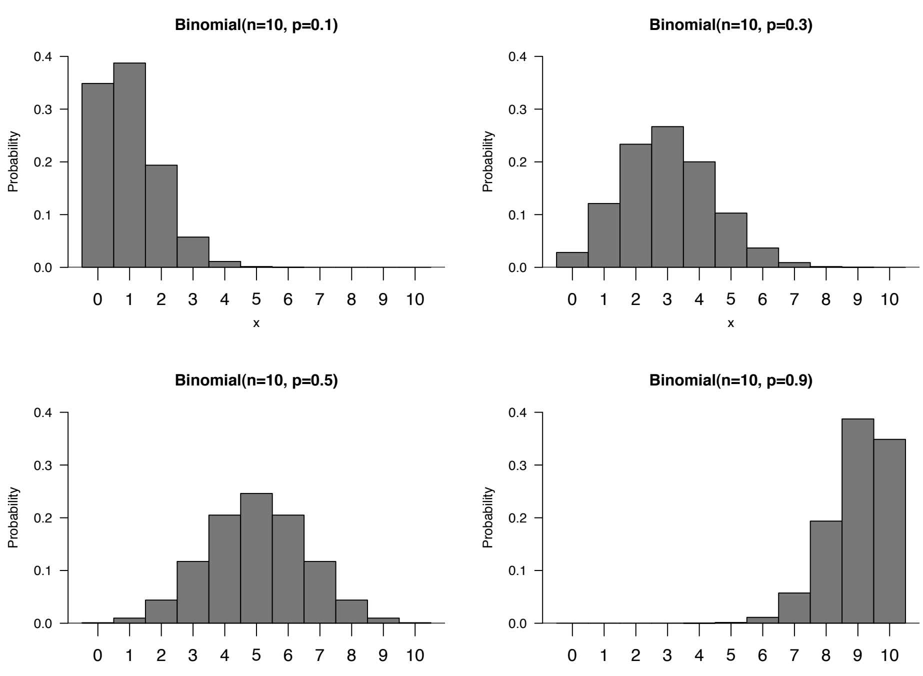
$\longrightarrow$ There are 13,983,816 possible combinations for the winning numbers.

Hence Pr({1,2,3,4,5,6}) = Pr({8,23,24,34,42,45}) = 1/13983816

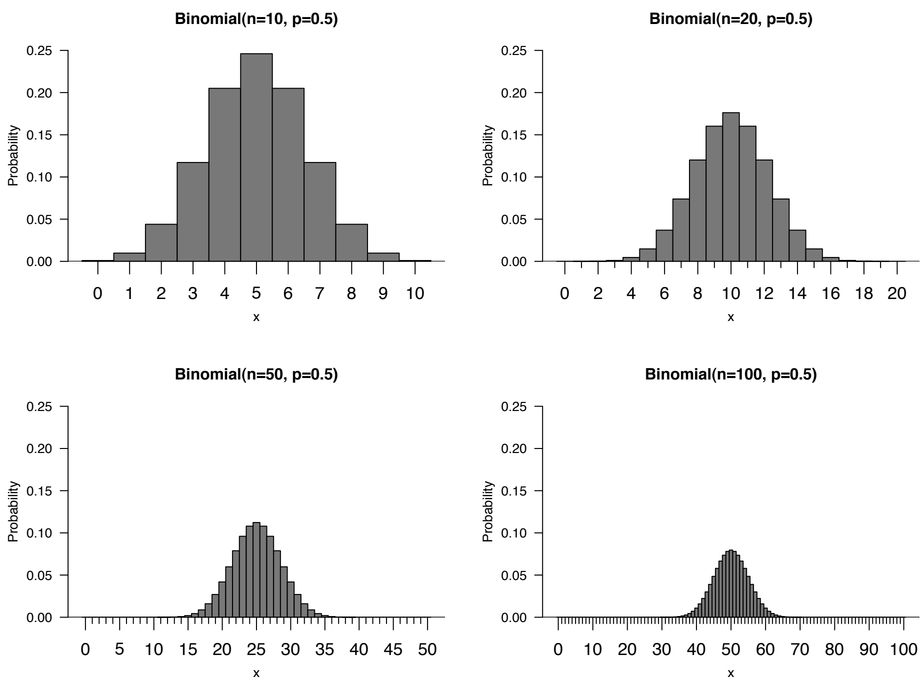
The probability of the having six consecutive numbers as the winning numbers is

$$\Pr(\{1,2,3,4,5,6\}) \;+\; \cdots \;+\; \Pr(\{44,45,46,47,48,49\})$$
$$= \; 44 \times (1/13983816) \;\approx\; 0.0003\%.$$

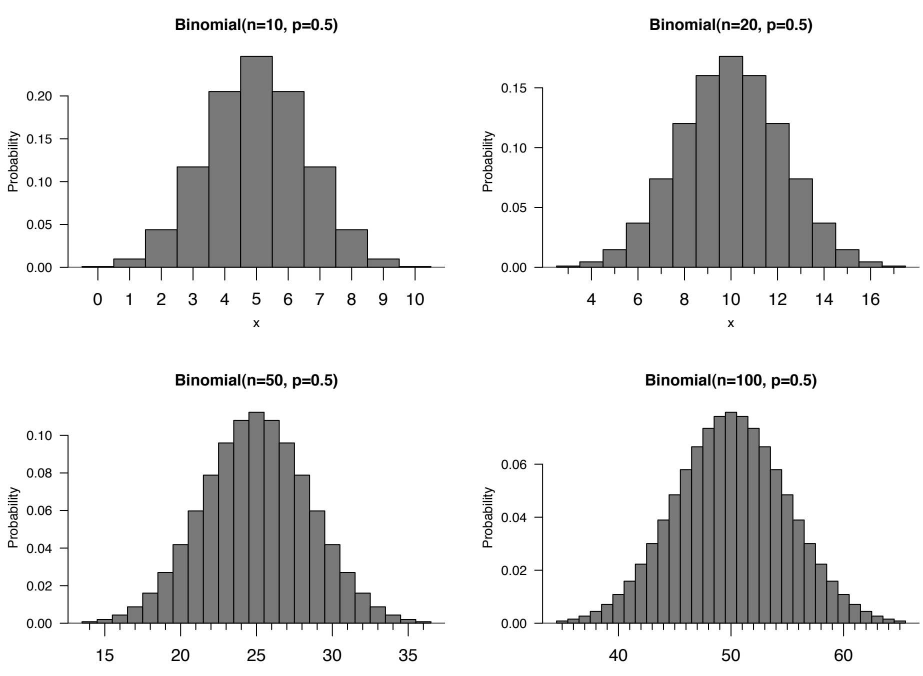
Binomial distributions



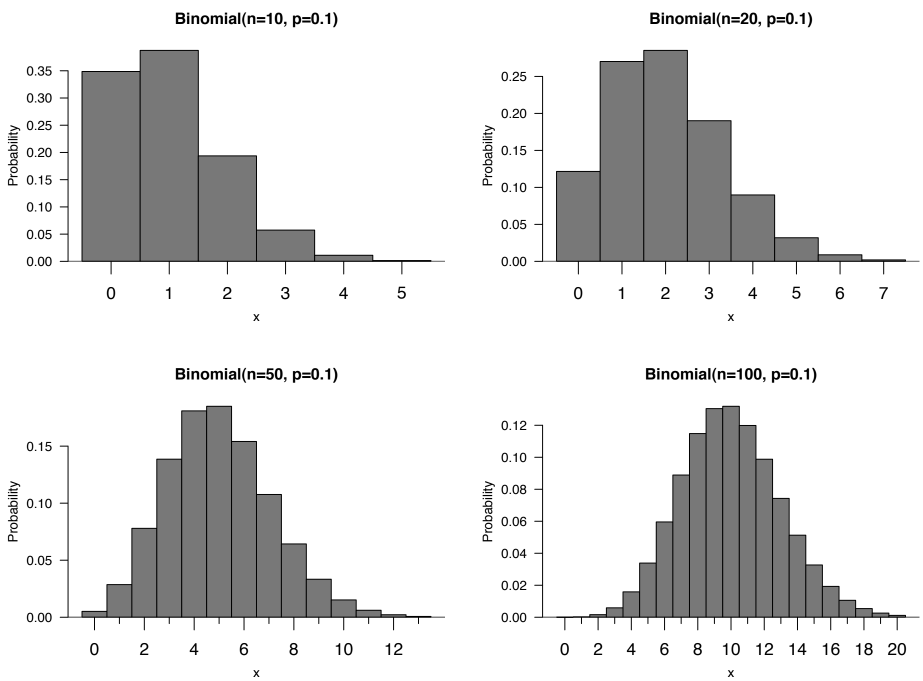
Binomial distributions



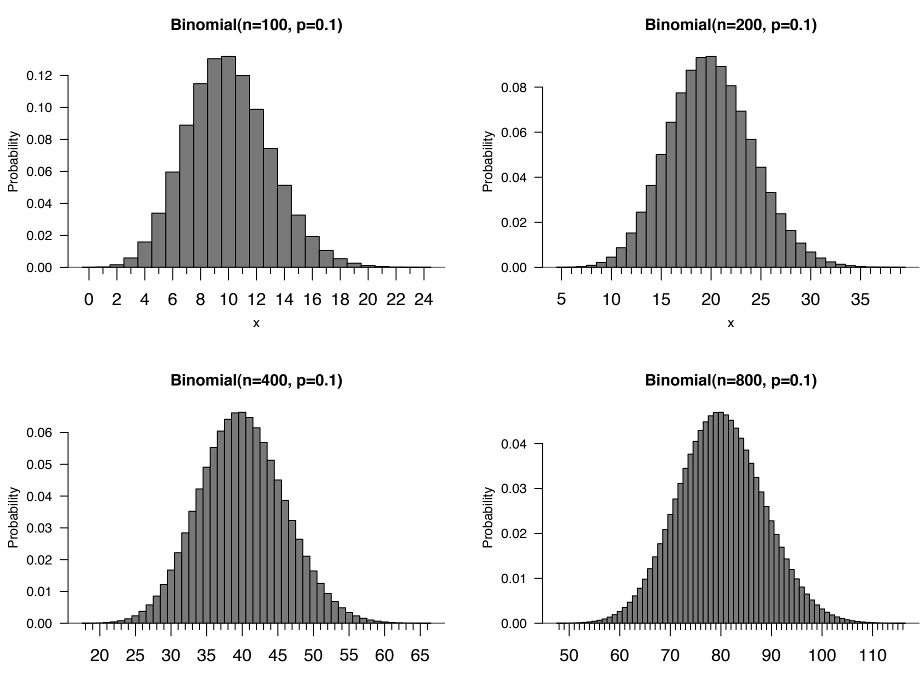
Binomial distributions



Binomial distributions



Binomial distributions



Expected value and standard deviation

- The expected value (or mean) of a discrete random variable X with probability function p(x) is

$$\mu = E(X) = \sum_x x\, p(x)$$

- The variance of a discrete random variable X with probability function p(x) is

$$\sigma^2 = \mathrm{var}(X) = \sum_x (x - \mu)^2\, p(x)$$

- The standard deviation (SD) of X is

$$\mathrm{SD}(X) = \sqrt{\mathrm{var}(X)}.$$

Mean and SD of binomial RVs

If $X \sim$ Binomial(n,p), then

$$E(X) = n\, p$$

$$\mathrm{SD}(X) = \sqrt{n\, p\, (1 - p)}$$

- Examples:

n	p	mean	SD
10	10%	1	0.9
10	30%	3	1.4
10	50%	5	1.6
10	90%	9	0.9

Binomial random variable

Number of successes in n trials where:

$\longrightarrow$ Trials independent

$\longrightarrow$ p = Pr(success) is constant

The number of successes in n trials does not necessarily follow a binomial distribution!

Deviations from the binomial:

$\longrightarrow$ Varying p

$\longrightarrow$ Clumping or repulsion (non-independence)

Examples

Suppose treatment response differs between genders:

Pr(responds | male) = 10%   but   Pr(responds | female) = 80%.

$\longrightarrow$ Pick 4 male and 6 female subjects.

The number of responders is not binomial.

$\longrightarrow$ Pick 10 random subjects (with Pr(subject is male) = 40%).

The number of responders is binomial.

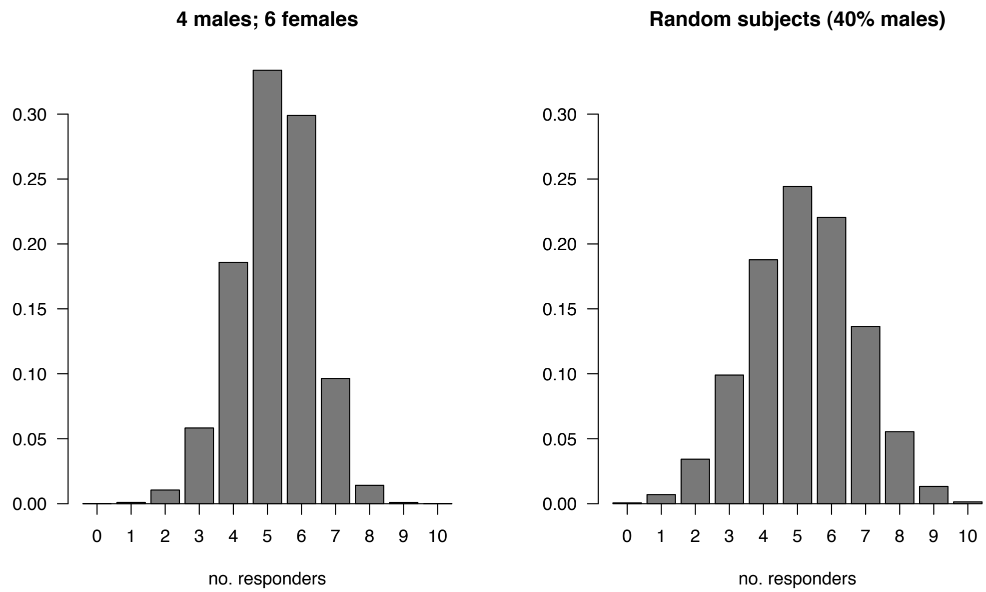
$p = 0.4 \times 0.1 + 0.6 \times 0.8 = 0.52$.

Pr(responds) =

Pr(responds and male) + Pr(responds and female) =

Pr(male) $\times$ Pr(responds | male) + Pr(female) $\times$ Pr(responds | female)

Examples



Poisson distribution

Consider a Binomial(n,p) where

$\longrightarrow$ n is really large

$\longrightarrow$ p is really small

For example, suppose each well in a microtiter plate contains 50,000 T cells, and that 1/100,000 cells respond to a particular antigen.

Let X be the number of responding cells in a well.

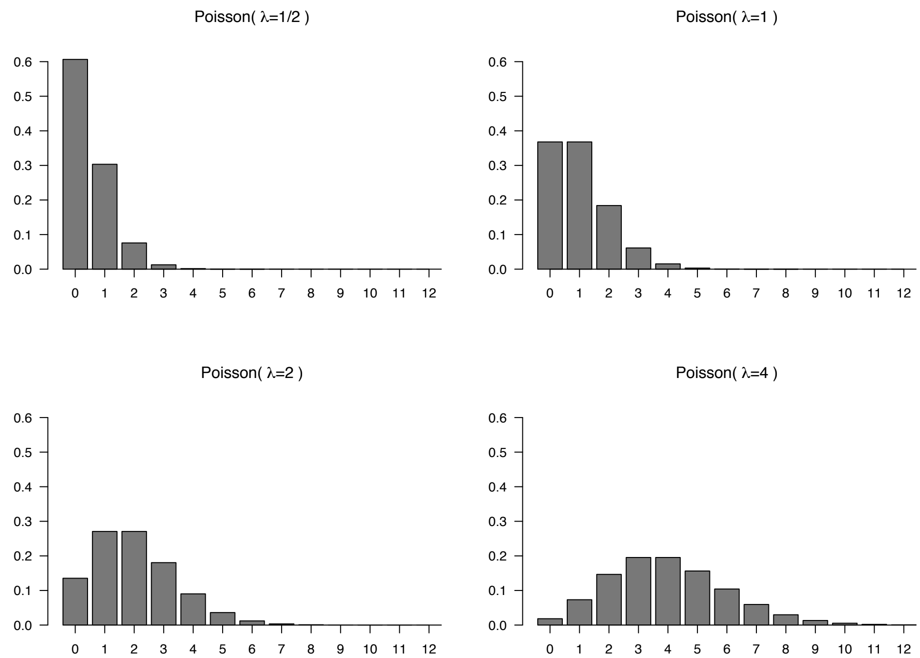
$\longrightarrow$ In this case, X follows a Poisson distribution approximately.

Let λ = n p = E(X).

$\longrightarrow$ p(x) = Pr(X = x) = $e^{-\lambda}\lambda^x/x!$

Note that SD(X) = $\sqrt{\lambda}$.

Poisson distribution



Example

Suppose there are 100,000 T cells in each well of a microtiter plate. Suppose that 1/80,000 T cells respond to a particular antigen.

Let X = number of responding T cells in a well.

$\longrightarrow$ $X \sim \text{Poisson}(\lambda = 1.25)$.

$\longrightarrow$ $\text{E}(X) = 1.25$

$\longrightarrow$ $\text{SD}(X) = \sqrt{1.25} \approx 1.12$.

$\Pr(X = 0) = \exp(-1.25) \approx 29\%$.

$\Pr(X > 0) = 1 - \exp(-1.25) \approx 71\%$.

$\Pr(X = 2) = \exp(-1.25) \times (1.25)^2 / 2 \approx 22\%$.

$Y = a + b\ X$

Suppose X is a discrete random variable with probability function p, so that p(x) = Pr(X = x).

$\longrightarrow$ Expected value: E(X) = $\sum_x$ x p(x)

$\longrightarrow$ Standard deviation: SD(X) = $\sqrt{\sum_x [x - E(X)]^2\ p(x)}$

Let Y = a + b X where a and b are numbers. Then Y is a random variable (like X), and

$\longrightarrow$ E(Y) = a + b E(X)

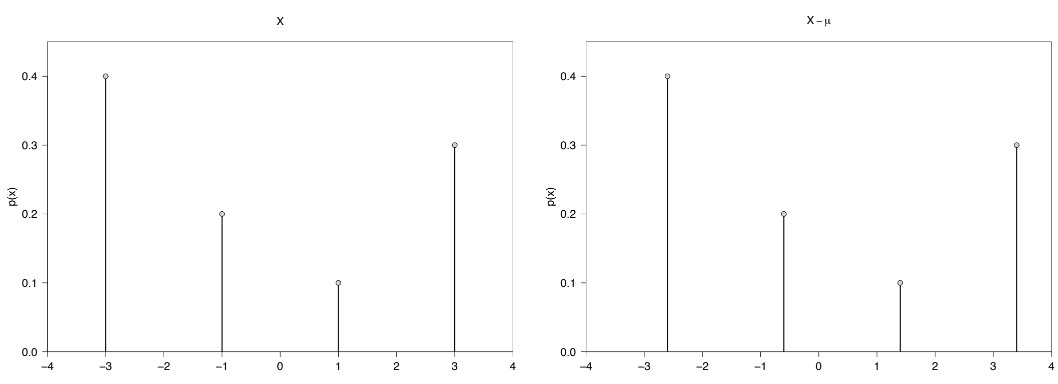
$\longrightarrow$ SD(Y) = |b| SD(X)

In particular, if μ = E(X), σ = SD(X), and Z = ($X - \mu$) / σ, then

$\longrightarrow$ E(Z) = 0

$\longrightarrow$ SD(Z) = 1

$Y = a + b\ X$



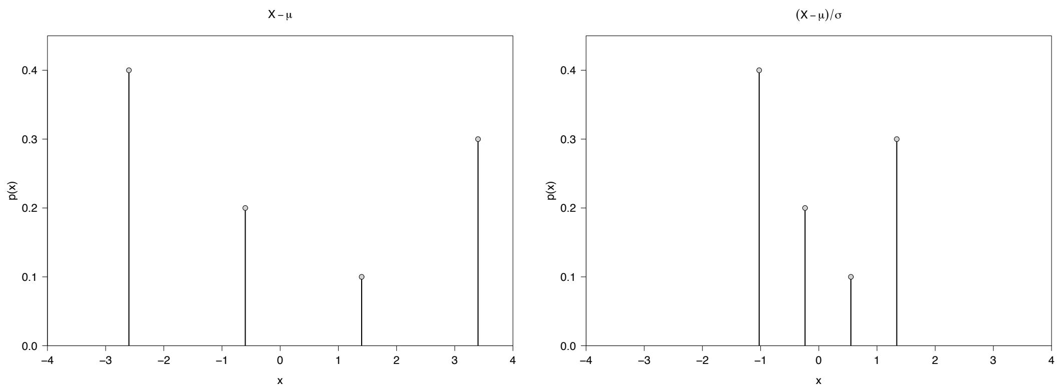
Let X be a random variable with mean μ and SD σ.

If $Y = X - \mu$, then

$\longrightarrow$ E(Y) = 0

$\longrightarrow$ SD(Y) = σ

$Y = a + b\,X$



Let X be a random variable with mean μ and SD σ.

If $Y = (X - \mu) / \sigma$, then

$\longrightarrow$ E(Y) = 0

$\longrightarrow$ SD(Y) = 1

Example

Suppose $X \sim$ Binomial(n,p) $\rightarrow$ number of successes

$\longrightarrow$ E(X) = n p

$\longrightarrow$ SD(X) = $\sqrt{\text{n p (1 – p)}}$

Let $P = X$ / n $\rightarrow$ proportion of successes

$\longrightarrow$ E(P) = E(X / n) = E(X) / n = p

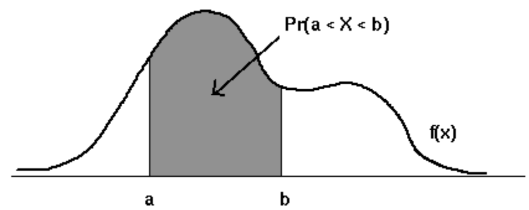
$\longrightarrow$ SD(P) = SD(X / n) = SD(X) / n = $\sqrt{\text{p (1 – p) / n}}$

Continuous random variables

Suppose X is a continuous random variable.

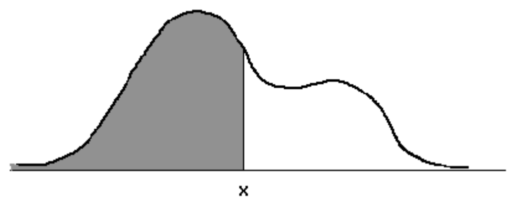
Instead of a probability function, X has a probability density function (pdf), sometimes called just the density of X.

- $\longrightarrow$ $f(x) \geq 0$
- $\longrightarrow$ $\int_{-\infty}^{\infty} f(x)\, d(x) = 1$
- $\longrightarrow$ Areas under curve = probabilities



Cumulative distr. function:

- $\longrightarrow$ $F(x) = Pr(X \leq x)$ $\longrightarrow$



Means and standard deviations

Expected value:

- $\longrightarrow$ Discrete RV: $E(X) = \sum_x x\, p(x)$
- $\longrightarrow$ Continuous RV: $E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx$

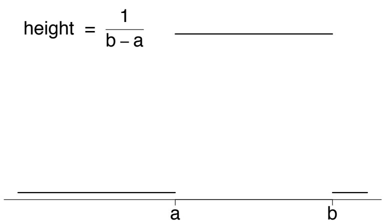
Standard deviation:

- $\longrightarrow$ Discrete RV: $SD(X) = \sqrt{\sum_x [x - E(X)]^2\, p(x)}$
- $\longrightarrow$ Continuous RV: $SD(X) = \sqrt{\int_{-\infty}^{\infty} [x - E(X)]^2\, f(x)\, dx}$

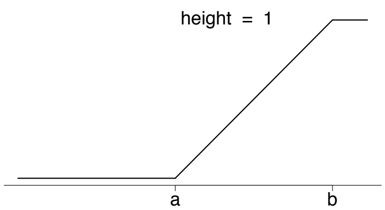
Uniform distribution

$X \sim$ Uniform(a, b)

$\longrightarrow$ Draw a number at random from the interval (a, b).



$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}$$



$\longrightarrow$ E(X) = (a + b) / 2

$\longrightarrow$ SD(X) = (b – a) / $\sqrt{12}$

$\approx 0.29 \times$ (b – a)

Normal distribution

By far the most important distribution:
The normal distribution (also called the Gaussian distribution).

If $X \sim$ N(μ, σ), then the pdf of X is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

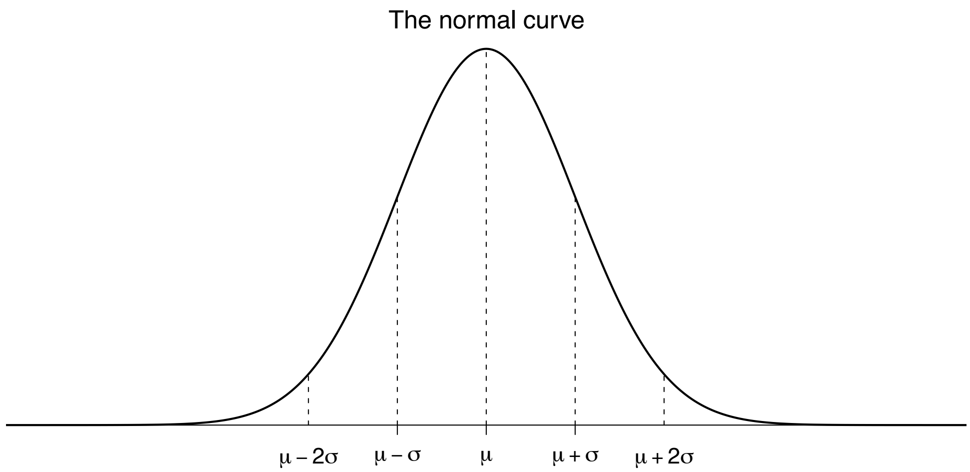
Note: E(X) = μ and SD(X) = σ.

Of great importance:

$\longrightarrow$ If $X \sim$ N(μ,σ) and Z = (X – μ) / σ, then $Z \sim$ N(0, 1).

This is the standard normal distribution.

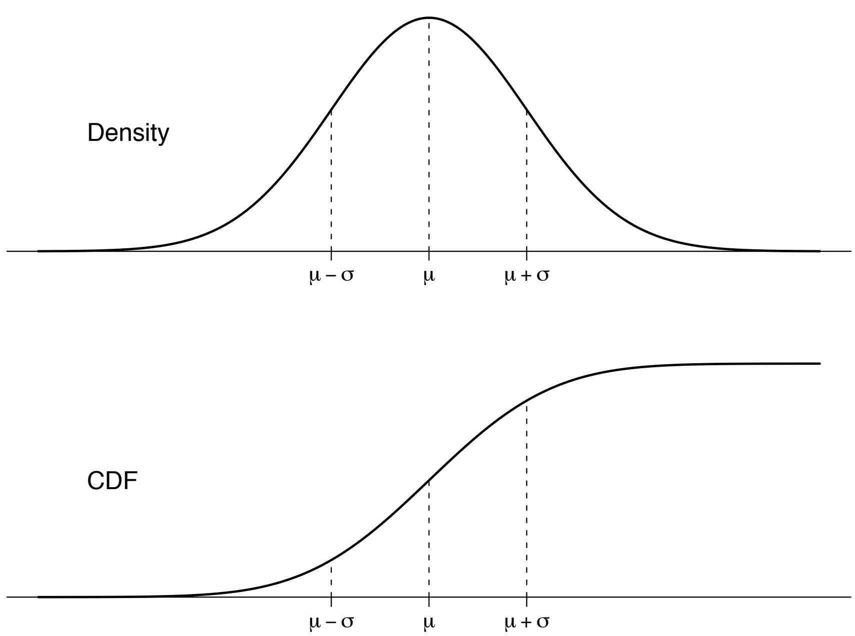
Normal distribution



⟶ Remember:

$\Pr(\mu - \sigma \leq X \leq \mu + \sigma) \approx 68\%$ and $\Pr(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 95\%$.

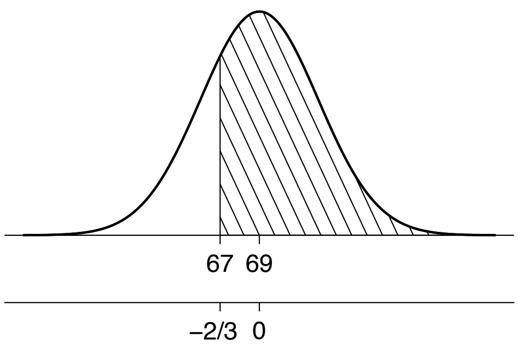
The normal CDF



Example

Suppose the heights of adult males in the U.S. are approximately normal distributed, with mean = 69 in and SD = 3 in.

$\longrightarrow$ What proportion of men are taller than 5'7"?



$X \sim N(\mu{=}69,\ \sigma{=}3)$

$Z = (X - 69)/3 \sim N(0,1)$

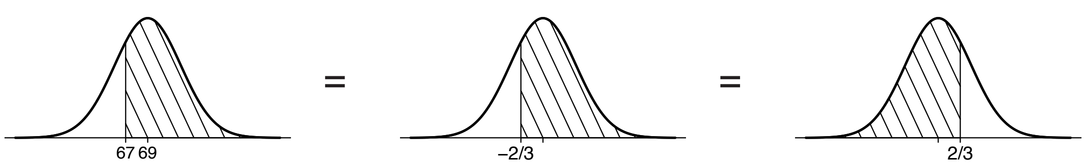
$\Pr(X \geq 67) =$

$\Pr(Z \geq (67 - 69)/3) =$

$\Pr(Z \geq -2/3)$

Example

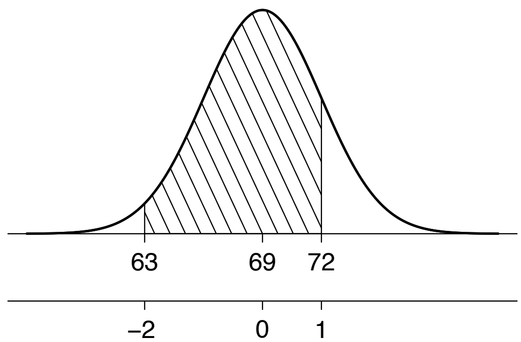
Use either of the following three:



The answer: 75%.

Another calculation

⟶ What proportion of men are between 5’3” and 6’?



$$\Pr(63 \leq X \leq 72) \;=\; \Pr(-2 \leq Z \leq 1) \;\longrightarrow\; 82\%.$$