Chapter 3: Distributions of Random Variables

OpenIntro Statistics, 3rd Edition

Slides developed by Mine Çetinkaya-Rundel of OpenIntro.

Random variables

Random variables

- A *random variable* is a numeric quantity whose value depends on the outcome of a random event
  - We use a capital letter, like X, to denote a random variable
  - The values of a random variable are denoted with a lowercase letter, in this case x
  - For example, $P(X = x)$
- There are two types of random variables:
  - *Discrete random variables* often take only integer values
    - Example: Number of credit hours, Difference in number of credit hours this term vs last
  - *Continuous random variables* take real (decimal) values
    - Example: Cost of books this term, Difference in cost of books this term vs last

Expectation

- We are often interested in the average outcome of a random variable.
- We call this the *expected value* (mean), and it is a weighted average of the possible outcomes

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

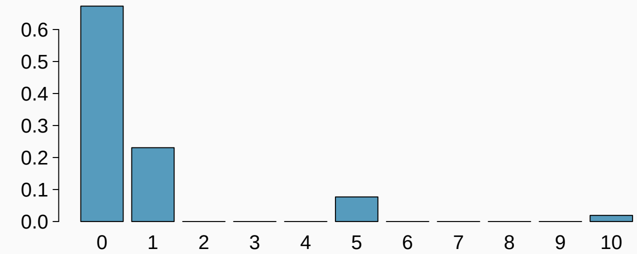
Expected value of a discrete random variable

In a game of cards you win \$1 if you draw a heart, \$5 if you draw an ace (including the ace of hearts), \$10 if you draw the king of spades and nothing for any other card you draw. Write the probability model for your winnings, and calculate your expected winning.

Event	X	$P(X)$	$X\,P(X)$
Heart (not ace)	1	$\frac{12}{52}$	$\frac{12}{52}$
Ace	5	$\frac{4}{52}$	$\frac{20}{52}$
King of spades	10	$\frac{1}{52}$	$\frac{10}{52}$
All else	0	$\frac{35}{52}$	0
Total			$E(X) = \frac{42}{52} \approx 0.81$

Expected value of a discrete random variable (cont.)

Below is a visual representation of the probability distribution of winnings from this game:



Variability

We are also often interested in the variability in the values of a random variable.

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

$$\sigma = SD(X) = \sqrt{Var(X)}$$

Variability of a discrete random variable

For the previous card game example, how much would you expect the winnings to vary from game to game?

X	$P(X)$	$X\,P(X)$	$(X - E(X))^2$	$P(X)\,(X - E(X))^2$
1	$\frac{12}{52}$	$1 \times \frac{12}{52} = \frac{12}{52}$	$(1 - 0.81)^2 = 0.0361$	$\frac{12}{52} \times 0.0361 = 0.0083$
5	$\frac{4}{52}$	$5 \times \frac{4}{52} = \frac{20}{52}$	$(5 - 0.81)^2 = 17.5561$	$\frac{4}{52} \times 17.5561 = 1.3505$
10	$\frac{1}{52}$	$10 \times \frac{1}{52} = \frac{10}{52}$	$(10 - 0.81)^2 = 84.4561$	$\frac{1}{52} \times 84.0889 = 1.6242$
0	$\frac{35}{52}$	$0 \times \frac{35}{52} = 0$	$(0 - 0.81)^2 = 0.6561$	$\frac{35}{52} \times 0.6561 = 0.4416$
		$E(X) = 0.81$		

Linear combinations

- A *linear combination* of random variables X and Y is given by

$$aX + bY$$

where a and b are some fixed numbers.
- The average value of a linear combination of random variables is given by

$$E(aX + bY) = a \times E(X) + b \times E(Y)$$

Calculating the expectation of a linear combination

On average you take 10 minutes for each statistics homework problem and 15 minutes for each chemistry homework problem. This week you have 5 statistics and 4 chemistry homework problems assigned. What is the total time you expect to spend on statistics and physics homework for the week?

Linear combinations

- The variability of a linear combination of two independent random variables is calculated as

$$V(aX + bY) = a^2 \times V(X) + b^2 \times V(Y)$$

- The standard deviation of the linear combination is the square root of the variance.

Note: If the random variables are not independent, the variance calculation gets a little more complicated and is beyond the scope of this course.

Calculating the variance of a linear combination

The standard deviation of the time you take for each statistics homework problem is 1.5 minutes, and it is 2 minutes for each chemistry problem. What is the standard deviation of the time you expect to spend on statistics and physics homework for the week if you have 5 statistics and 4 chemistry homework problems assigned?

Practice

A casino game costs \$5 to play. If the first card you draw is red, then you get to draw a second card (without replacement). If the second card is the ace of clubs, you win \$500. If not, you don't win anything, i.e. lose your \$5. What is your expected profits/losses from playing this game? Remember: profit/loss = winnings - cost.

(a) A profit of 5¢

(b) A loss of 10¢

(c) A loss of 25¢

(d) A loss of 30¢

Fair game

A *fair* game is defined as a game that costs as much as its expected payout, i.e. expected profit is 0.

Do you think casino games in Vegas cost more or less than their expected payouts?

Simplifying random variables

Random variables do not work like normal algebraic variables:

$$X + X \neq 2X$$

$E(X+X) = E(X) + E(X) = 2E(X)$ $\quad$ $Var(X+X) = Var(X) + Var(X)$ (assuming independence) $= 2\ Var(X)$

$E(2X) = 2E(X)$ $\quad$ $Var(2X) = 2^2\ Var(X) = 4\ Var(X)$

$E(X+X) = E(2X)$, but $Var(X+X) \neq Var(2X)$.

Adding or multiplying?

A company has 5 Lincoln Town Cars in its fleet. Historical data show that annual maintenance cost for each car is on average \$2,154 with a standard deviation of \$132. What is the mean and the standard deviation of the total annual maintenance cost for this fleet?

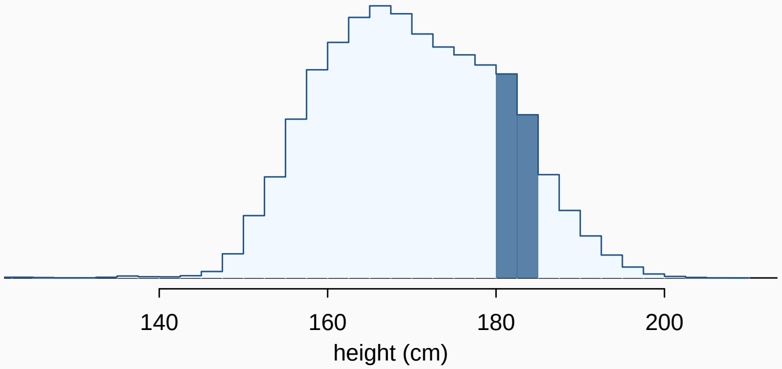
Note that we have 5 cars each with the given annual maintenance cost ($X_1 + X_2 + X_3 + X_4 + X_5$), not one car that had 5 times the given annual maintenance cost ($5X$).

$$\begin{aligned}
E(X_1 + X_2 + X_3 + X_4 + X_5) &= E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5) \\
&= 5 \times E(X) = 5 \times 2,154 = \$10,770 \\
Var(X_1 + X_2 + X_3 + X_4 + X_5) &= Var(X_1) + Var(X_2) + Var(X_3) + Var(X_4) + Var \\
&= 5 \times V(X) = 5 \times 132^2 = \$87,120 \\
SD(X_1 + X_2 + X_3 + X_4 + X_5) &= \sqrt{87,120} = 295.16
\end{aligned}$$

Continuous distributions

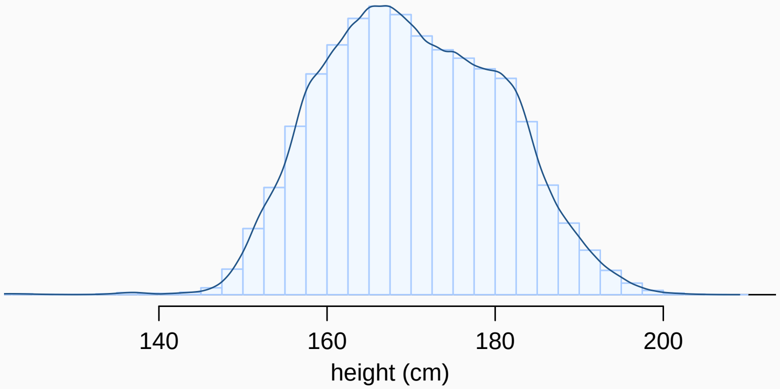
Continuous distributions

- Below is a histogram of the distribution of heights of US adults.
- The proportion of data that falls in the shaded bins gives the probability that a randomly sampled US adult is between 180 cm and 185 cm (about 5'11" to 6'1").



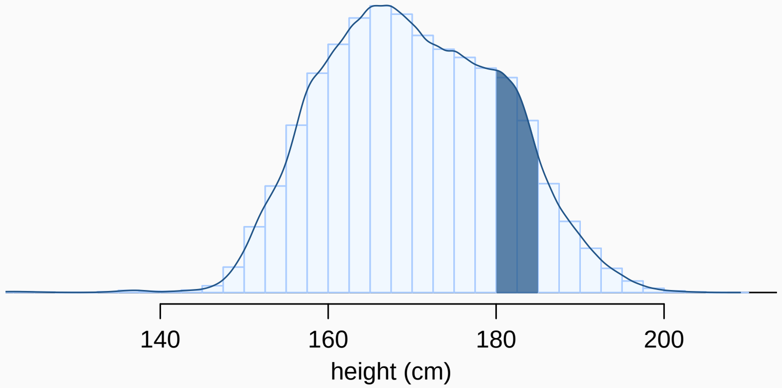
From histograms to continuous distributions

Since height is a continuous numerical variable, its *probability density function* is a smooth curve.



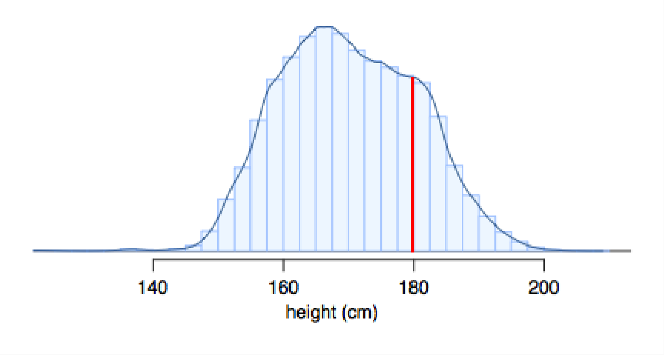
Probabilities from continuous distributions

Therefore, the probability that a randomly sampled US adult is between 180 cm and 185 cm can also be estimated as the shaded area under the curve.



By definition...

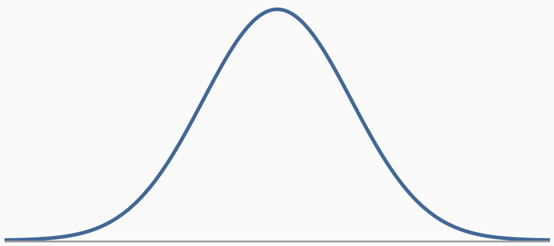
Since continuous probabilities are estimated as “the area under the curve”, the probability of a person being exactly 180 cm (or any exact value) is defined as 0.



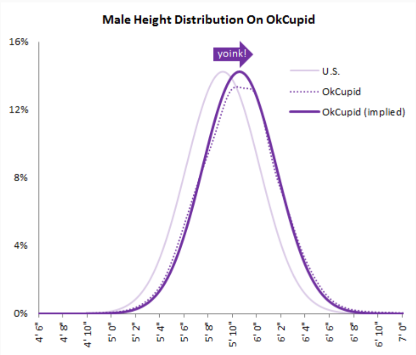
Normal distribution

Normal distribution

- Unimodal and symmetric, bell shaped curve
- Many variables are nearly normal, but none are exactly normal
- Denoted as $N(\mu, \sigma) \rightarrow$ Normal with mean μ and standard deviation σ



Heights of males

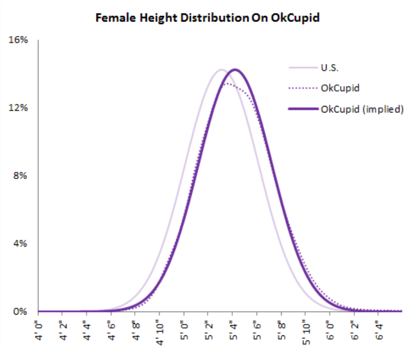


“The male heights on OkCupid very nearly follow the expected normal distribution – except the whole thing is shifted to the right of where it should be. Almost universally guys like to add a couple inches.”

“You can also see a more subtle vanity at work: starting at roughly 5’ 8”, the top of the dotted curve tilts even further rightward. This means that guys as they get closer to six feet round up a bit more than usual, stretching for that coveted psychological benchmark.”

http://blog.okcupid.com/index.php/the-biggest-lies-in-online-dating/

Heights of females



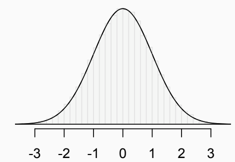
"When we looked into the data for women, we were surprised to see height exaggeration was just as widespread, though without the lurch towards a benchmark height."

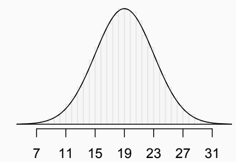
http://blog.okcupid.com/index.php/the-biggest-lies-in-online-dating/

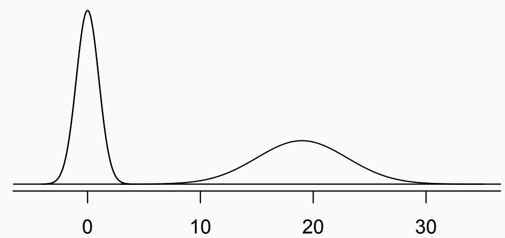
Normal distributions with different parameters

μ: mean, σ: standard deviation

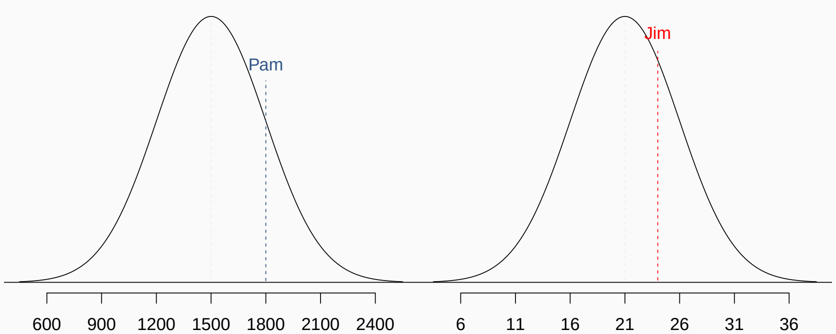
$N(\mu = 0, \sigma = 1)$ $N(\mu = 19, \sigma = 4)$







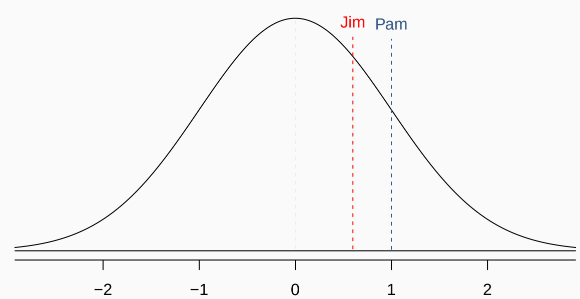
SAT scores are distributed nearly normally with mean 1500 and standard deviation 300. ACT scores are distributed nearly normally with mean 21 and standard deviation 5. A college admissions officer wants to determine which of the two applicants scored better on their standardized test with respect to the other test takers: Pam, who earned an 1800 on her SAT, or Jim, who scored a 24 on his ACT?



Standardizing with Z scores

Since we cannot just compare these two raw scores, we instead compare how many standard deviations beyond the mean each observation is.

- Pam's score is $\frac{1800-1500}{300} = 1$ standard deviation above the mean.
- Jim's score is $\frac{24-21}{5} = 0.6$ standard deviations above the mean.



Standardizing with Z scores (cont.)

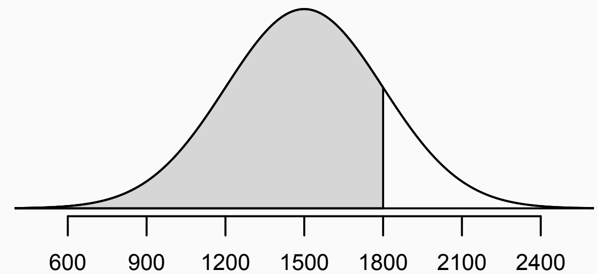
- These are called *standardized* scores, or *Z scores*.
- Z score of an observation is the number of standard deviations it falls above or below the mean.

$$Z = \frac{observation - mean}{SD}$$

- Z scores are defined for distributions of any shape, but only when the distribution is normal can we use Z scores to calculate percentiles.
- Observations that are more than 2 SD away from the mean ($|Z| > 2$) are usually considered unusual.

Percentiles

- *Percentile* is the percentage of observations that fall below a given data point.
- Graphically, percentile is the area below the probability distribution curve to the left of that observation.



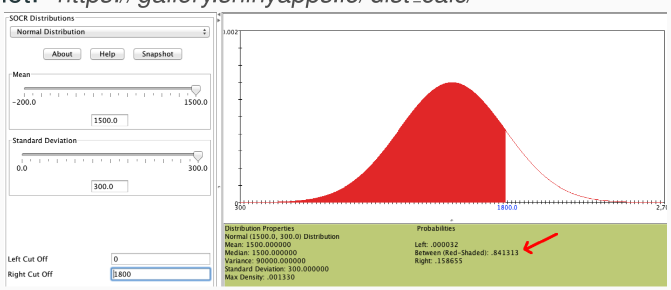
Calculating percentiles - using computation

There are many ways to compute percentiles/areas under the curve:

- R:

```
> pnorm(1800, mean = 1500, sd = 300)
[1] 0.8413447
```

- Applet: *https://gallery.shinyapps.io/dist_calc/*



Calculating percentiles - using tables

	Second decimal place of Z									
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	*0.8413*	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015

Six sigma

"The term *six sigma process* comes from the notion that if one has six standard deviations between the process mean and the nearest specification limit, as shown in the graph, practically no items will fail to meet specifications."

$$6\sigma$$

http://en.wikipedia.org/wiki/Six_Sigma

Quality control

At Heinz ketchup factory the amounts which go into bottles of ketchup are supposed to be normally distributed with mean 36 oz. and standard deviation 0.11 oz. Once every 30 minutes a bottle is selected from the production line, and its contents are noted precisely. If the amount of ketchup in the bottle is below 35.8 oz. or above 36.2 oz., then the bottle fails the quality control inspection. What percent of bottles have less than 35.8 ounces of ketchup?

Finding the exact probability - using the Z table

Second decimal place of Z										
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	Z
0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019	−2.9
0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026	−2.8
0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035	−2.7
0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047	−2.6
0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062	−2.5
0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082	−2.4
0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107	−2.3
0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139	−2.2
0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179	−2.1
0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228	−2.0
0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287	−1.9
0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359	−1.8
0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446	−1.7
0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548	−1.6
0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668	−1.5

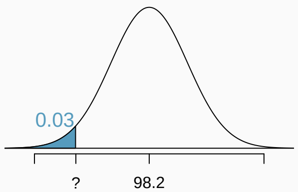
Practice

What percent of bottles <u>pass</u> the quality control inspection?

(a) 1.82%
(b) 3.44%
(c) 6.88%
(d) 93.12%
(e) 96.56%

Finding cutoff points

Body temperatures of healthy humans are distributed nearly normally with mean 98.2°F and standard deviation 0.73°F. What is the cutoff for the lowest 3% of human body temperatures?



0.09	0.08	0.07	0.06	0.05	Z
0.0233	0.0239	0.0244	0.0250	0.0256	−1.9
0.0294	0.0301	0.0307	0.0314	0.0322	−1.8
0.0367	0.0375	0.0384	0.0392	0.0401	−1.7

$$\begin{aligned} P(X < x) &= 0.03 \rightarrow P(Z < -1.88) = 0.03 \\ Z &= \frac{obs - mean}{SD} \rightarrow \frac{x - 98.2}{0.73} = -1.88 \\ x &= (-1.88 \times 0.73) + 98.2 = 96.8^{\circ}F \end{aligned}$$

Mackowiak, Wasserman, and Levine (1992), *A Critical Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlick.*

Practice

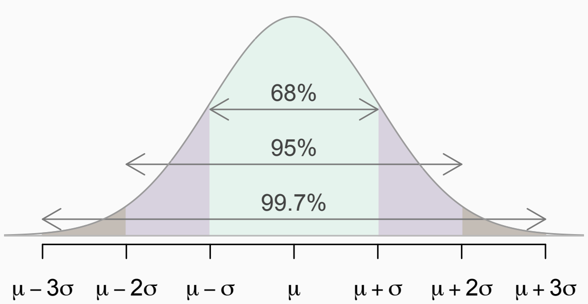
Body temperatures of healthy humans are distributed nearly normally with mean 98.2°F and standard deviation 0.73°F. What is the cutoff for the highest 10% of human body temperatures?

(a) 97.3°F
(b) 99.1°F
(c) 99.4°F
(d) 99.6°F

$$\begin{aligned} P(X > x) &= 0.10 \rightarrow P(Z < 1.28) = 0.90 \\ Z &= \frac{obs - mean}{SD} \rightarrow \frac{x - 98.2}{0.73} = 1.28 \\ x &= (1.28 \times 0.73) + 98.2 = 99.1 \end{aligned}$$

68-95-99.7 Rule

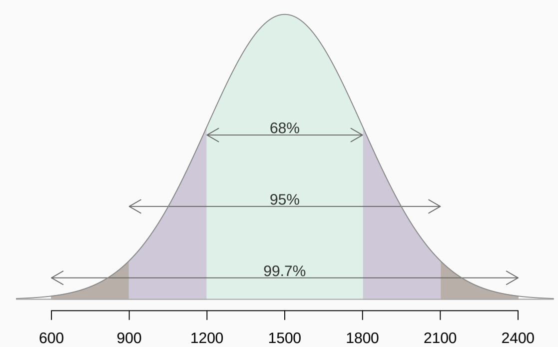
- For nearly normally distributed data,
  - about 68% falls within 1 SD of the mean,
  - about 95% falls within 2 SD of the mean,
  - about 99.7% falls within 3 SD of the mean.
- It is possible for observations to fall 4, 5, or more standard deviations away from the mean, but these occurrences are very rare if the data are nearly normal.



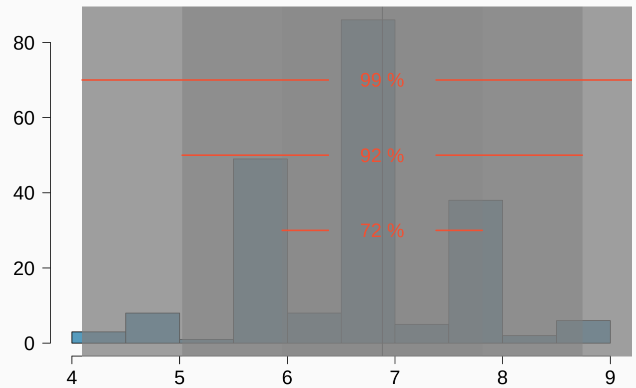
Describing variability using the 68-95-99.7 Rule

SAT scores are distributed nearly normally with mean 1500 and standard deviation 300.

- ~68% of students score between 1200 and 1800 on the SAT.
- ~95% of students score between 900 and 2100 on the SAT.
- ~99.7% of students score between 600 and 2400 on the SAT.



Number of hours of sleep on school nights



- Mean = 6.88 hours, SD = 0.92 hrs
- 72% of the data are within 1 SD of the mean: 6.88 ± 0.93
- 92% of the data are within 1 SD of the mean: $6.88 \pm 2 \times 0.93$
- 99% of the data are within 1 SD of the mean: $6.88 \pm 3 \times 0.93$

Practice

Which of the following is <u>false</u>?

(a) Majority of Z scores in a right skewed distribution are negative.

(b) In skewed distributions the Z score of the mean might be different than 0.

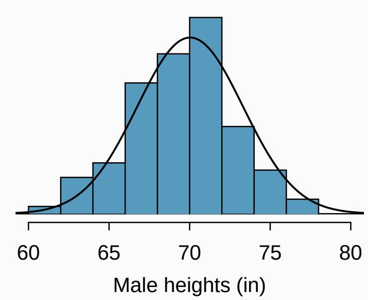
(c) For a normal distribution, IQR is less than $2 \times SD$.

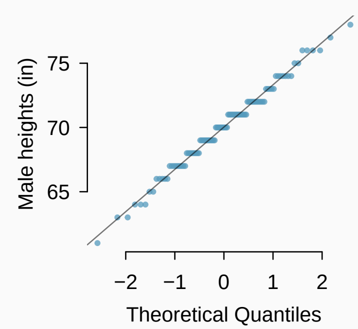
(d) Z scores are helpful for determining how unusual a data point is compared to the rest of the data in the distribution.

Evaluating the normal approximation

Normal probability plot

A histogram and *normal probability plot* of a sample of 100 male heights.

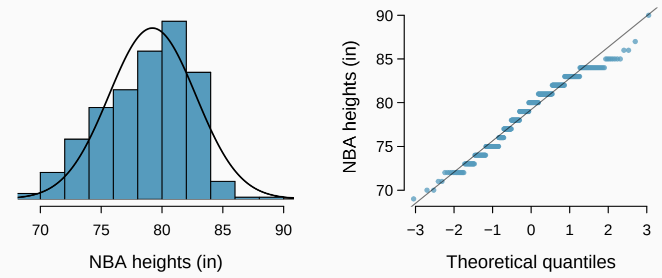




Anatomy of a normal probability plot

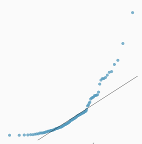
- Data are plotted on the y-axis of a normal probability plot, and theoretical quantiles (following a normal distribution) on the x-axis.
- If there is a linear relationship in the plot, then the data follow a nearly normal distribution.
- Constructing a normal probability plot requires calculating percentiles and corresponding z-scores for each observation, which is tedious. Therefore we generally rely on software when making these plots.

Below is a histogram and normal probability plot for the NBA heights from the 2008-2009 season. Do these data appear to follow a normal distribution?

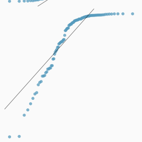


Why do the points on the normal probability have jumps?

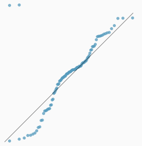
Normal probability plot and skewness



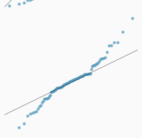
Right skew - Points bend up and to the left of the line.



Left skew- Points bend down and to the right of the line.



Short tails (narrower than the normal distribution) - Points follow an S shaped-curve.

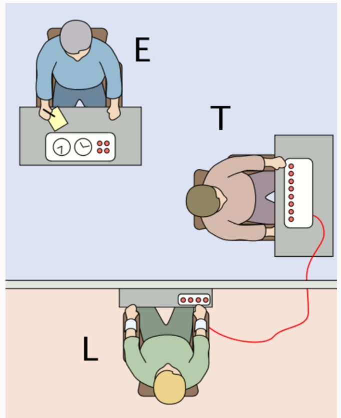


Long tails (wider than the normal distribution) - Points start below the line, bend to follow it, and end above it.

Geometric distribution

Milgram experiment

- Stanley Milgram, a Yale University psychologist, conducted a series of experiments on obedience to authority starting in 1963.
- Experimenter (E) orders the teacher (T), the subject of the experiment, to give severe electric shocks to a learner (L) each time the learner answers a question incorrectly.
- The learner is actually an actor, and the electric shocks are not real, but a prerecorded sound is played each time the teacher administers an electric shock.



http://en.wikipedia.org/wiki/File:Milgram_Experiment_v2.png

Milgram experiment (cont.)

- These experiments measured the willingness of study participants to obey an authority figure who instructed them to perform acts that conflicted with their personal conscience.
- Milgram found that about 65% of people would obey authority and give such shocks.
- Over the years, additional research suggested this number is approximately consistent across communities and time.

Bernouilli random variables

- Each person in Milgram's experiment can be thought of as a *trial*.
- A person is labeled a *success* if she refuses to administer a severe shock, and *failure* if she administers such shock.
- Since only 35% of people refused to administer a shock, *probability of success* is $p = 0.35$.
- When an individual trial has only two possible outcomes, it is called a *Bernoulli random variable*.

Geometric distribution

Dr. Smith wants to repeat Milgram's experiments but she only wants to sample people until she finds someone who will not inflict a severe shock. What is the probability that she stops after the first person?

$$P(1^{st} \text{ person refuses}) = 0.35$$

... the third person?

$$P(1^{st} \text{ and } 2^{nd} \text{ shock}, 3^{rd} \text{ refuses}) = \underset{0.65}{\underline{S}} \times \underset{0.65}{\underline{S}} \times \underset{0.35}{\underline{R}} = 0.65^2 \times 0.35 \approx 0.15$$

... the tenth person?

Geometric distribution (cont.)

Geometric distribution describes the waiting time until a success for *independent and identically distributed (iid)* Bernouilli random variables.

- independence: outcomes of trials don't affect each other
- identical: the probability of success is the same for each trial

Geometric probabilities

If p represents probability of success, $(1 - p)$ represents probability of failure, and n represents number of independent trials

$$P(\textit{success on the } n^{th} \textit{ trial}) = (1 - p)^{n-1}p$$

Can we calculate the probability of rolling a 6 for the first time on the 6^{th} roll of a die using the geometric distribution? Note that what was a success (rolling a 6) and what was a failure (not rolling a 6) are clearly defined and one or the other must happen for each trial.

(a) no, on the roll of a die there are more than 2 possible outcomes

(b) yes, why not

Expected value

How many people is Dr. Smith expected to test before finding the first one that refuses to administer the shock?

The expected value, or the mean, of a geometric distribution is defined as $\frac{1}{p}$.

$$\mu = \frac{1}{p} = \frac{1}{0.35} = 2.86$$

She is expected to test 2.86 people before finding the first one that refuses to administer the shock.

But how can she test a non-whole number of people?

Expected value and its variability

Mean and standard deviation of geometric distribution

$$\mu = \frac{1}{p} \qquad \sigma = \sqrt{\frac{1-p}{p^2}}$$

- Going back to Dr. Smith's experiment:

$$\sigma = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1-0.35}{0.35^2}} = 2.3$$

- Dr. Smith is expected to test 2.86 people before finding the first one that refuses to administer the shock, give or take 2.3 people.
- These values only make sense in the context of repeating the experiment many many times.

Binomial distribution

Suppose we randomly select four individuals to participate in this experiment. What is the probability that exactly 1 of them will refuse to administer the shock?

Let's call these people Allen (A), Brittany (B), Caroline (C), and Damian (D). Each one of the four scenarios below will satisfy the condition of "exactly 1 of them refuses to administer the shock":

$$\text{Scenario 1:} \quad \underset{\text{(A) } \textit{refuse}}{\underline{0.35}} \times \underset{\text{(B) shock}}{\underline{0.65}} \times \underset{\text{(C) shock}}{\underline{0.65}} \times \underset{\text{(D) shock}}{\underline{0.65}} = 0.0961$$

$$\text{Scenario 2:} \quad \underset{\text{(A) shock}}{\underline{0.65}} \times \underset{\text{(B) } \textit{refuse}}{\underline{0.35}} \times \underset{\text{(C) shock}}{\underline{0.65}} \times \underset{\text{(D) shock}}{\underline{0.65}} = 0.0961$$

$$\text{Scenario 3:} \quad \underset{\text{(A) shock}}{\underline{0.65}} \times \underset{\text{(B) shock}}{\underline{0.65}} \times \underset{\text{(C) } \textit{refuse}}{\underline{0.35}} \times \underset{\text{(D) shock}}{\underline{0.65}} = 0.0961$$

$$\text{Scenario 4:} \quad \underset{\text{(A) shock}}{\underline{0.65}} \times \underset{\text{(B) shock}}{\underline{0.65}} \times \underset{\text{(C) shock}}{\underline{0.65}} \times \underset{\text{(D) } \textit{refuse}}{\underline{0.35}} = 0.0961$$

The probability of exactly one 1 of 4 people refusing to administer the shock is the sum of all of these probabilities.

$$0.0961 + 0.0961 + 0.0961 + 0.0961 = 4 \times 0.0961 = 0.3844$$

Binomial distribution

The question from the prior slide asked for the probability of given number of successes, k, in a given number of trials, n, ($k = 1$ success in $n = 4$ trials), and we calculated this probability as

$$\# \textit{ of scenarios} \times P(\textit{single scenario})$$

- $\#$ *of scenarios*: there is a less tedious way to figure this out, we'll get to that shortly...
- $P(\textit{single scenario}) = p^k \ (1 - p)^{(n-k)}$

  probability of success to the power of number of successes, probability of failure to the power of number of failures

The *Binomial distribution* describes the probability of having exactly k successes in n independent Bernouilli trials with probability of success p.

Counting the # of scenarios

Earlier we wrote out all possible scenarios that fit the condition of exactly one person refusing to administer the shock. If n was larger and/or k was different than 1, for example, $n = 9$ and $k = 2$:

*RR*SSSSSSS
S*RR*SSSSSS
SS*RR*SSSSS
...
SS*R*SS*R*SSS
...
SSSSSSS*RR*

writing out all possible scenarios would be incredibly tedious and prone to errors.

Calculating the # of scenarios

Choose function

The *choose function* is useful for calculating the number of ways to choose k successes in n trials.

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

- $k = 1$, $n = 4$: $\binom{4}{1} = \frac{4!}{1!(4-1)!} = \frac{4 \times 3 \times 2 \times 1}{1 \times (3 \times 2 \times 1)} = 4$
- $k = 2$, $n = 9$: $\binom{9}{2} = \frac{9!}{2!(9-1)!} = \frac{9 \times 8 \times 7!}{2 \times 1 \times 7!} = \frac{72}{2} = 36$

Note: You can also use R for these calculations:

```
> choose(9,2)
[1] 36
```


Properties of the choose function

Which of the following is false?

(a) There are n ways of getting 1 success in n trials, $\binom{n}{1} = n$.

(b) There is only 1 way of getting n successes in n trials, $\binom{n}{n} = 1$.

(c) There is only 1 way of getting n failures in n trials, $\binom{n}{0} = 1$.

(d) There are $n-1$ ways of getting $n-1$ successes in n trials, $\binom{n}{n-1} = n - 1$.

Binomial distribution (cont.)

Binomial probabilities

If p represents probability of success, $(1 - p)$ represents probability of failure, n represents number of independent trials, and k represents number of successes

$$P(k \text{ successes in } n \text{ trials}) = \binom{n}{k} p^k \, (1 - p)^{(n-k)}$$

Which of the following is not a condition that needs to be met for the binomial distribution to be applicable?

(a) the trials must be independent

(b) the number of trials, n, must be fixed

(c) each trial outcome must be classified as a *success* or a *failure*

(d) the number of desired successes, k, must be greater than the number of trials

(e) the probability of success, p, must be the same for each trial

A 2012 Gallup survey suggests that 26.2% of Americans are obese. Among a random sample of 10 Americans, what is the probability that exactly 8 are obese?

(a) pretty high

(b) pretty low

Gallup: *http://www.gallup.com/poll/160061/obesity-rate-stable-2012.aspx*, January 23, 2013.

A 2012 Gallup survey suggests that 26.2% of Americans are obese. Among a random sample of 10 Americans, what is the probability that exactly 8 are obese?

(a) $0.262^8 \times 0.738^2$

(b) $\binom{8}{10} \times 0.262^8 \times 0.738^2$

(c) $\binom{10}{8} \times 0.262^8 \times 0.738^2$

(d) $\binom{10}{8} \times 0.262^2 \times 0.738^8$

The birthday problem

What is the probability that 2 randomly chosen people share a birthday?

Pretty low, $\frac{1}{365} \approx 0.0027$.

What is the probability that at least 2 people out of 366 people share a birthday?

Exactly 1! (Excluding the possibility of a leap year birthday.)

The birthday problem (cont.)

What is the probability that at least 2 people (1 match) out of 121 people share a birthday?

Somewhat complicated to calculate, but we can think of it as the complement of the probability that there are no matches in 121 people.

$$
\begin{aligned}
P(\textit{no matches}) &= 1 \times \left(1 - \frac{1}{365}\right) \times \left(1 - \frac{2}{365}\right) \times \cdots \times \left(1 - \frac{120}{365}\right) \\
&= \frac{365 \times 364 \times \cdots \times 245}{365^{121}} \\
&= \frac{365!}{365^{121} \times (365 - 121)!} \\
&= \frac{121! \times \binom{365}{121}}{365^{121}} \approx 0 \\
P(\textit{at least 1 match}) &\approx 1
\end{aligned}
$$

Expected value

A 2012 Gallup survey suggests that 26.2% of Americans are obese.

Among a random sample of 100 Americans, how many would you expect to be obese?

- Easy enough, $100 \times 0.262 = 26.2$.
- Or more formally, $\mu = np = 100 \times 0.262 = 26.2$.
- But this doesn't mean in every random sample of 100 people exactly 26.2 will be obese. In fact, that's not even possible. In some samples this value will be less, and in others more. How much would we expect this value to vary?

Expected value and its variability

Mean and standard deviation of binomial distribution

$$\mu = np \qquad \sigma = \sqrt{np(1-p)}$$

- Going back to the obesity rate:

$$\sigma = \sqrt{np(1-p)} = \sqrt{100 \times 0.262 \times 0.738} \approx 4.4$$

- We would expect 26.2 out of 100 randomly sampled Americans to be obese, with a standard deviation of 4.4.

Note: Mean and standard deviation of a binomial might not always be whole numbers, and that is alright, these values represent what we would expect to see on average.

Unusual observations

Using the notion that *observations that are more than 2 standard deviations away from the mean are considered unusual* and the mean and the standard deviation we just computed, we can calculate a range for the plausible number of obese Americans in random samples of 100.

$$26.2 \pm (2 \times 4.4) = (17.4, 35)$$

An August 2012 Gallup poll suggests that 13% of Americans think home schooling provides an excellent education for children. Would a random sample of 1,000 Americans where only 100 share this opinion be considered unusual?

(a) No (b) Yes

	Excellent	Good	Only fair	Poor	Total excellent/ good
	%	%	%	%	%
Independent private school	31	47	13	2	78
Parochial or church-related schools	21	48	18	5	69
Charter schools	17	43	23	5	60
Home schooling	13	33	30	14	46
Public schools	5	32	42	19	37

Gallup, Aug. 9-12, 2012

http://www.gallup.com/poll/156974/private-schools-top-marks-educating-children.aspx

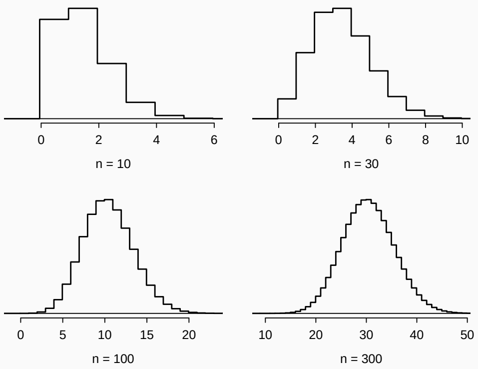
Shapes of binomial distributions

For this activity you will use a web applet. Go to *https://gallery.shinyapps.io/dist_calc/* and choose Binomial coin experiment in the drop down menu on the left.

- Set the number of trials to 20 and the probability of success to 0.15. Describe the shape of the distribution of number of successes.
- Keeping p constant at 0.15, determine the minimum sample size required to obtain a unimodal and symmetric distribution of number of successes. Please submit only one response per team.
- Further considerations:
  - What happens to the shape of the distribution as n stays constant and p changes?
  - What happens to the shape of the distribution as p stays constant and n changes?

Distributions of number of successes

Hollow histograms of samples from the binomial model where $p = 0.10$ and $n = 10, 30, 100,$ and 300. What happens as n increases?



Low large is large enough?

The sample size is considered large enough if the expected number of successes and failures are both at least 10.

$$np \geq 10 \qquad \text{and} \qquad n(1-p) \geq 10$$

Below are four pairs of Binomial distribution parameters. Which distribution can be approximated by the normal distribution?

(a) $n = 100, p = 0.95$

(b) $n = 25, p = 0.45$

(c) $n = 150, p = 0.05$

(d) $n = 500, p = 0.015$

An analysis of Facebook users

A recent study found that "Facebook users get more than they give". For example:

- 40% of Facebook users in our sample made a friend request, but 63% received at least one request
- Users in our sample pressed the like button next to friends' content an average of 14 times, but had their content "liked" an average of 20 times
- Users sent 9 personal messages, but received 12
- 12% of users tagged a friend in a photo, but 35% were themselves tagged in a photo

Any guesses for how this pattern can be explained?

http://www.pewinternet.org/Reports/2012/Facebook-users/Summary.aspx

This study also found that approximately 25% of Facebook users are considered power users. The same study found that the average Facebook user has 245 friends. What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users? Note any assumptions you must make.

We are given that $n = 245, p = 0.25$, and we are asked for the probability $P(K \geq 70)$. To proceed, we need independence, which we'll assume but could check if we had access to more Facebook data.

$$P(X \geq 70) = P(K = 70 \text{ or } K = 71 \text{ or } K = 72 \text{ or } \cdots \text{ or } K = 245)$$
$$= P(K = 70) + P(K = 71) + P(K = 72) + \cdots + P(K = 245)$$

This seems like an awful lot of work...

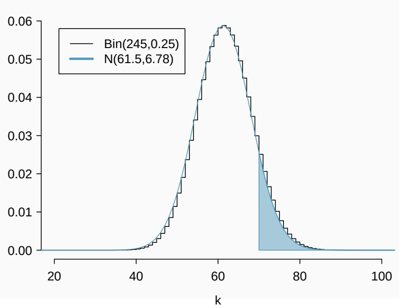
Normal approximation to the binomial

When the sample size is large enough, the binomial distribution with parameters n and p can be approximated by the normal model with parameters $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.

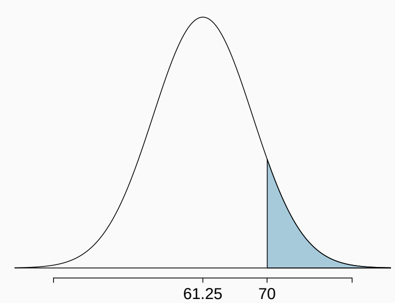
- In the case of the Facebook power users, $n = 245$ and $p = 0.25$.

  $$\mu = 245 \times 0.25 = 61.25 \qquad \sigma = \sqrt{245 \times 0.25 \times 0.75} = 6.78$$

- $Bin(n = 245, p = 0.25) \approx N(\mu = 61.25, \sigma = 6.78)$.



What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users?



$$Z = \frac{obs - mean}{SD} = \frac{70 - 61.25}{6.78} = 1.29$$

	Second decimal place of Z				
Z	0.05	0.06	0.07	0.08	0.09
1.0	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8944	0.8962	0.8980	0.8997	0.9015

$$P(Z > 1.29) = 1 - 0.9015 = 0.0985$$