OpenIntro Statistics

CH 03: Distributions of Random Variables

Models

- "All models are wrong, but some are useful."
  George Box, statistician

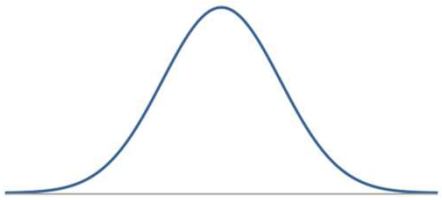
Normal distribution

Normal Distribution

Unimodal and symmetric, bell shaped curve

Many variables are nearly normal, but none are exactly normal

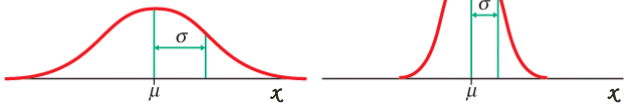
Denoted as $N(\mu, \sigma)$ → Normal with mean μ and standard deviation σ



Normal distributions

Normal—or Gaussian—distributions are a family of symmetrical, bell- shaped density curves defined by a mean μ (*mu*) and a standard deviation σ (*sigma*): $N(\mu, \sigma)$.

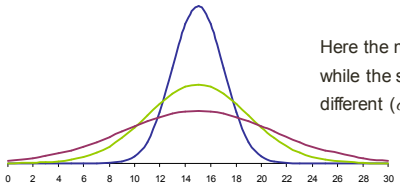
$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



e = 2.71828… The base of the natural logarithm
π = pi = 3.14159…

5

A family of density curves



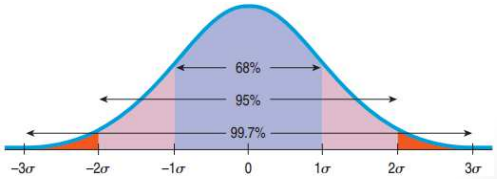
Here the means are the same ($\mu = 15$) while the standard deviations are different ($\sigma = 2$, 4, and 6).

Here the means are different ($\mu = 10$, 15, and 20) while the standard deviations are the same ($\sigma = 3$).

6

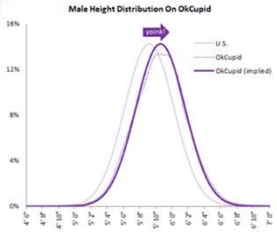
The 68-95-99.7 Rule

- 68% of the values fall within 1 standard deviation of the mean.
- 95% of the values fall within 2 standard deviations of the mean.
- 99.7% of the values fall within 3 standard deviations of the mean.



7

Heights of males



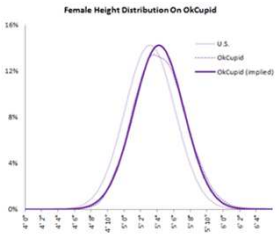
"The male heights on OkCupid very nearly follow the expected normal distribution -- except the whole thing is shifted to the right of where it should be. Almost universally guys like to add a couple inches."

"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

8

Heights of females

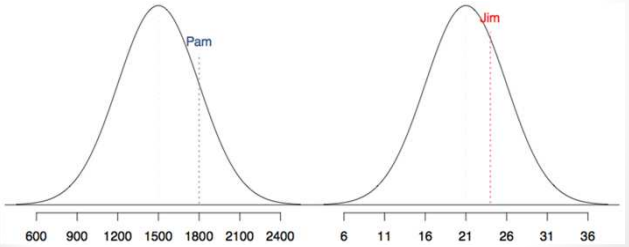


"When we looked into the data for women, we were surprised to see height exaggeration was just as widespread, though without the lurch towards a benchmark height."

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

9

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?

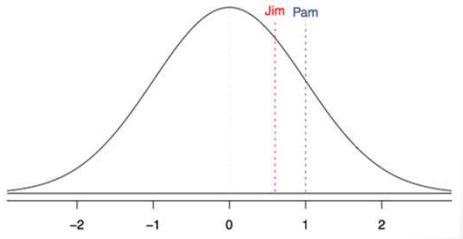


10

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 (1800 - 1500) / 300 = 1 standard deviation above the mean.
- Jim's score is (24 - 21) / 5 = 0.6 standard deviations above the mean.



11

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.

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

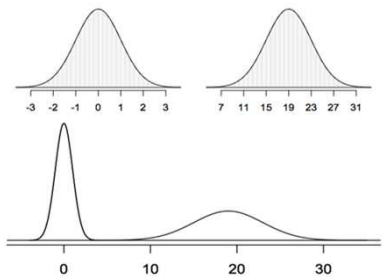
- Z scores are defined for distributions of any shape, but only when the distribution is normal can we use Z scores to calculate percentiles.
- Observations that are more than 2 SD away from the mean ($|Z| > 2$) are usually considered unusual.

12

Normal distributions with different parameters

μ: mean, σ: standard deviation

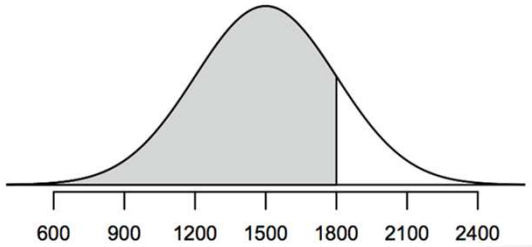
$N(\mu = 0, \sigma = 1)$ $\quad$ $N(\mu = 19, \sigma = 4)$



13

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.



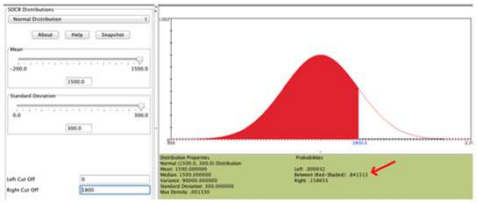
14

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: www.socr.ucla.edu/htmls/SOCR_Distributions.html



15

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

16

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.

$$6\sigma$$

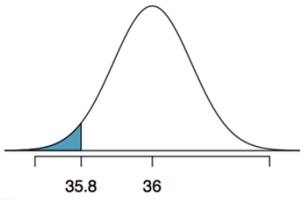
http://en.wikipedia.org/wiki/Six_Sigma

17

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?

- *Let X = amount of ketchup in a bottle: X ~ N(μ = 36, σ = 0.11)*



$$Z = \frac{35.8 - 36}{0.11} = -1.82$$

18

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

19

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

20

Practice

What percent of bottles pass the quality control inspection?

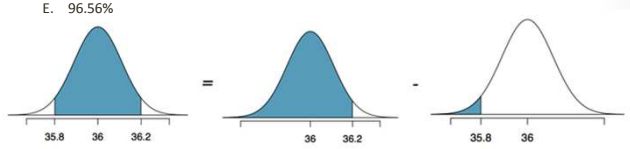
A. 1.82%
B. 3.44%
C. 6.88%
D. 93.12%
E. 96.56%

21

Practice

What percent of bottles pass the quality control inspection?

A. 1.82%
B. 3.44%
C. 6.88%
D. 93.12%
E. 96.56%



$$Z_{35.8} = \frac{35.8 - 36}{0.11} = -1.82$$

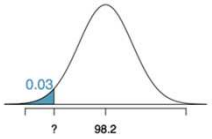
$$Z_{36.2} = \frac{36.2 - 36}{0.11} = 1.82$$

$$P(35.8 < X < 36.2) = P(-1.82 < Z < 1.82) = 0.9656 - 0.0344 = 0.9312$$

22

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

$$P(X < x) = 0.03 \rightarrow P(Z < -1.88) = 0.03$$

$$Z = \frac{obs - mean}{SD} \rightarrow \frac{x - 98.2}{0.73} = -1.88$$

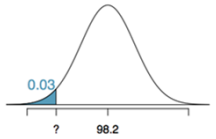
$$x = (-1.88 \times 0.73) + 98.2 = 96.8°F$$

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.

23

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?



We can do this easily in R:

```
> qnorm(p = 0.03,mean = 98.2,sd = 0.73)
[1] 96.82702
```

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.

24

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?

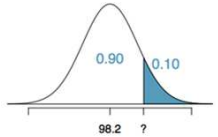
A. 97.3°F
B. 99.1°F
C. 99.4°F
D. 99.6°F

25

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




```
qnorm(p = 0.9,mean = 98.2,sd = 0.73)
[1] 99.13553
```

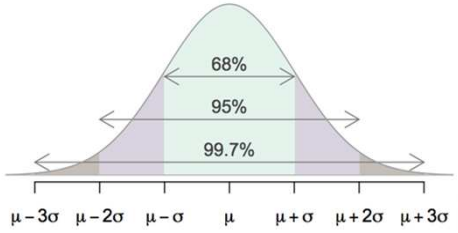
26

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.

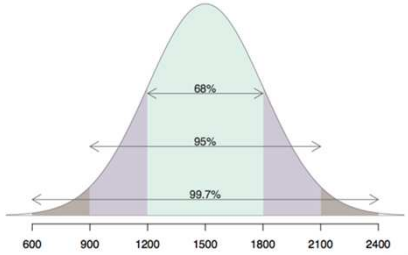


27

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.



28

Practice

Which of the following is false?

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 x SD.
D. Z scores are helpful for determining how unusual a data point is compared to the rest of the data in the distribution.

29

Practice

Which of the following is false?

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 x SD.
D. Z scores are helpful for determining how unusual a data point is compared to the rest of the data in the distribution.

30

Evaluating the Normal Approximation

There are two visual methods for checking the assumption of normality

- Histogram
- Normal Probability Plot

31

Evaluating the Normal Approximation

Normal Probability Plot

Plots each value against the *z*-score that would be expected had the distribution been perfectly normal.

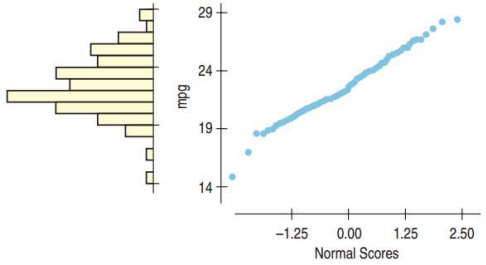
If the plot shows a line or is nearly straight, then the Normal model works.

If the plot strays from being a line, then the Normal model is not a good model.

Tends to curve at the tails.

32

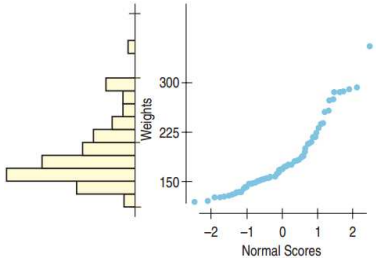
The Normal Model Applies



• The Normal probability plot is nearly straight, so the Normal model applies. Note that the histogram is unimodal and somewhat symmetric.

33

The Normal Model Does Not Apply



• The Normal probability plot is not straight, so the Normal model does not apply. Note that the histogram is skewed right.

34

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.

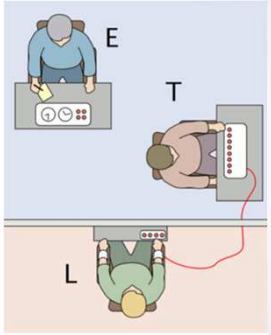
35

3.3 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 pre-recorded sound is played each time the teacher administers an electric shock.



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

37

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.

38

Bernoulli 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*.

39

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?

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

... the third person?

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

... the tenth person?

$$P(9\ shock,\ 10^{th}\ refuses) = \underbrace{\underset{0.65}{S} \times \cdots \times \underset{0.65}{S}}_{9\ of\ these} \times \underset{0.35}{R} = 0.65^9 \times 0.35 \approx 0.0072$$

40

Geometric distribution (cont.)

The *geometric distribution* describes the waiting time until a success for *independent and identically distributed (iid)* Bernoulli 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

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

41

Practice

Can we calculate the probability of rolling a 6 for the first time on the 6th 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

42

Practice

Can we calculate the probability of rolling a 6 for the first time on the 6th 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

$$P(6 \text{ on the } 6^{th} \text{ roll}) = \left(\frac{5}{6}\right)^5 \left(\frac{1}{6}\right) \approx 0.067$$

43

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 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?

44

Expected value and its variability

- Mean and standard deviation of geometric distribution

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

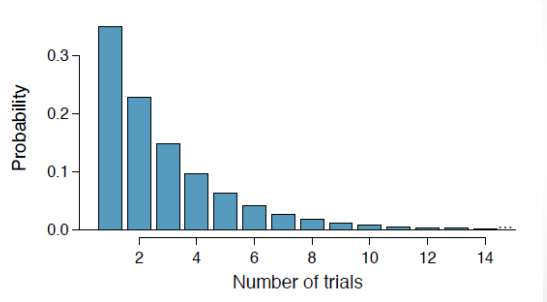
- Going back to Dr. Smith's experiment:

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

- 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.

45

Geometric Distribution, $p = 0.35$



46