Lecture 12 (and 14 and 15)

Some Notes

Lectures

These slides are **long** - we definitely won't finish today. These slides will be used for Lecture 14 (Wednesday after reading week) and 15 (Friday after reading week) as well.

Midterm

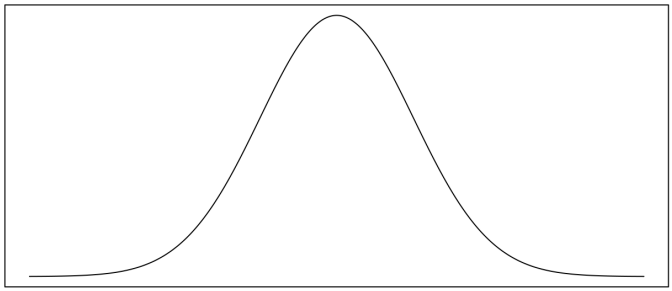
Reminder that our midterm is on Friday! Please show up on time, and bring at least two writing implements (pen or pencil is fine) and a hand calculator. It will make things easier if you don't bring a backpack - space will be an issue.

Further reminder: as you enter the room for the midterm, come to the front, and fill in the middle first. Do **not** sit on the ends of rows, or someone will snap at you for being in the way (that someone might be me …).

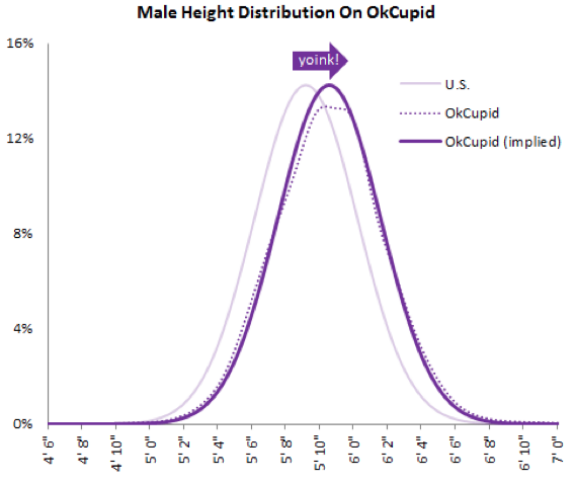
Normal Distribution

The Normal Distribution

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



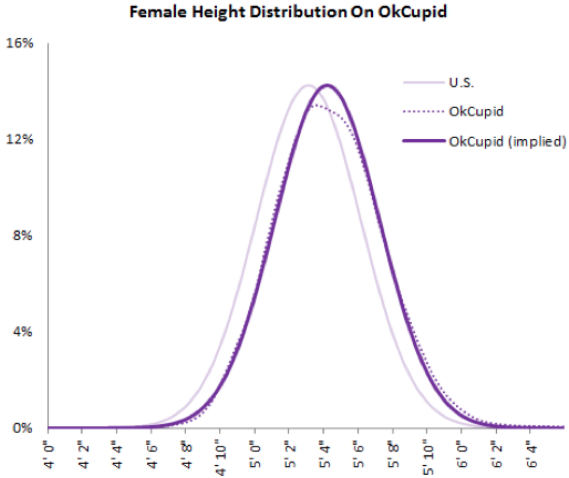
Heights of Males



"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."

Heights of Females

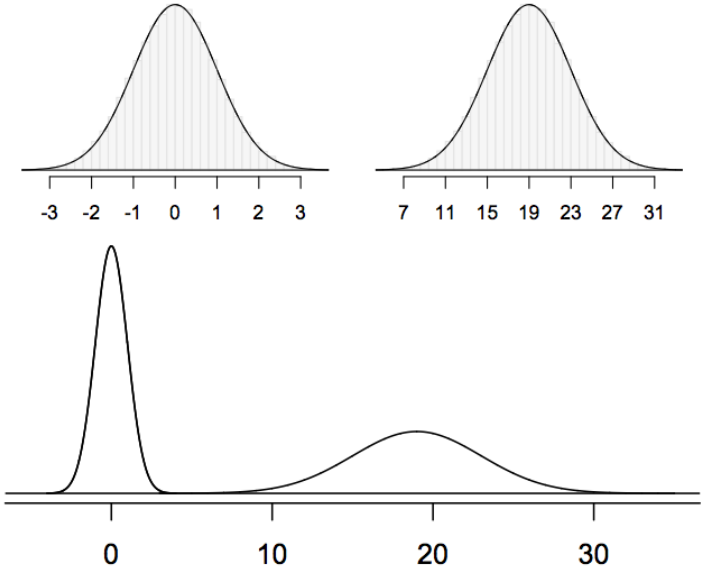


"When we looked into the data for women, we were surprised to see height exaggeration was just as widespread, though without the lurch towards a benchmark height."

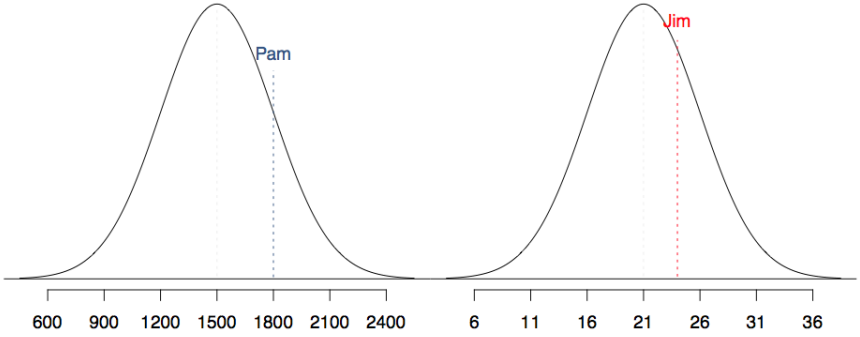
Normal distributions with different parameters

μ: mean, σ: standard deviation

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



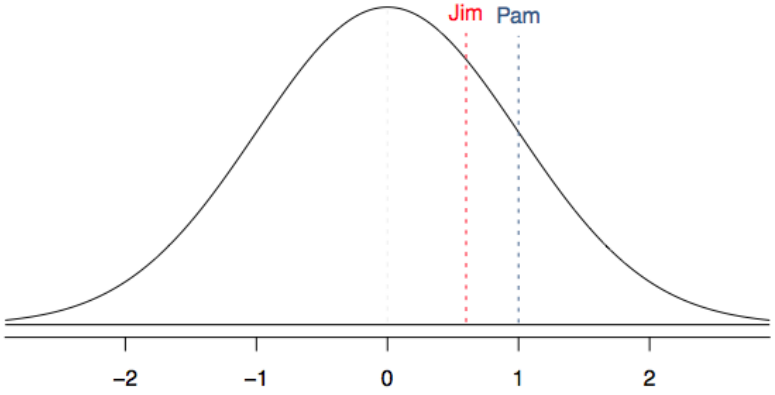
SAT scores are distributed nearly normally with mean 1500 and standard deviation 300. ACT scores are distributed nearly normally with mean 21 and standard deviation 5. A college admissions officer wants to determine which of the two applicants scored better on their standardized test with respect to 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 $(1800 - 1500)/300 = 1$ standard deviation above the mean.
- Jim's score is $(24 - 21)/5 = 0.6$ standard deviations above the mean.



Standardizing with Z-scores (continued)

These are called **standardized scores**, or **Z-scores** (or **Z scores**).

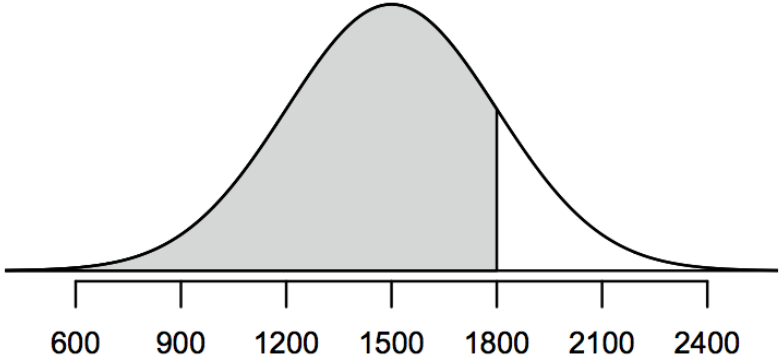
- Z score of an observation is the number of standard deviations it falls above or below the mean.

$$Z = (\text{observation} - \text{mean})/\text{SD}$$

- 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

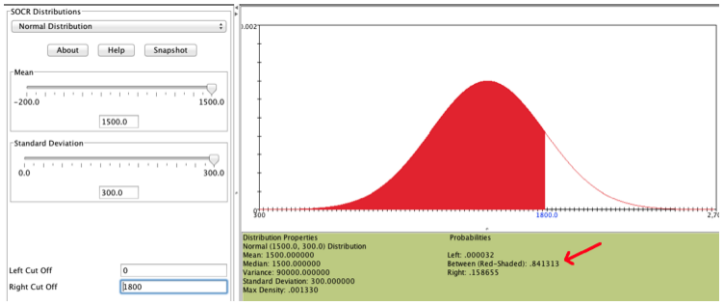
There are many ways to compute percentiles/areas under the curve.

R:

```
pnorm(1800, mean = 1500, sd = 300)
```

```
## [1] 0.8413447
```

Applets:



Calculating Percentiles - Look them Up!

	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

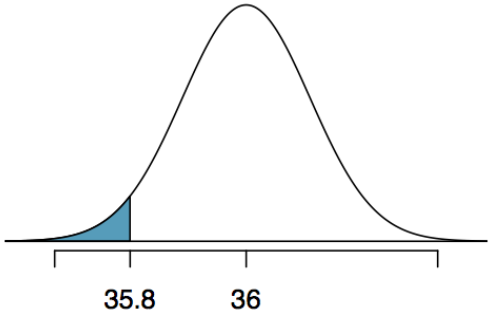
Six Sigma

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.

Example: Quality Control

At the 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: $X \sim \mathcal{N}(\mu = 36, \sigma = 0.11)$



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

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

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

Finding the exact probability - using R

```
pnorm(-1.82)
```

```
## [1] 0.0343795
```

Simpler!

Practice

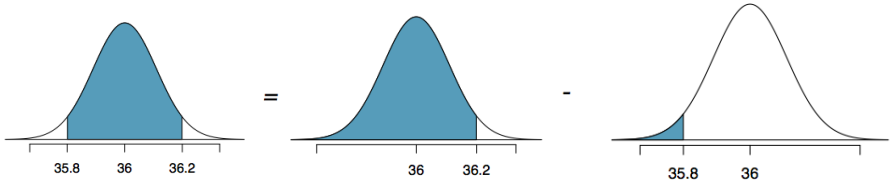
What percentage of bottles **pass** the quality control inspection?

1. 1.82%
2. 3.44%
3. 6.88%
4. 93.12%
5. 95.56%

Practice

What percentage of bottles **pass** the quality control inspection?

1. 1.82%
2. 3.44%
3. 6.88%
4. *93.12%*
5. 95.56%



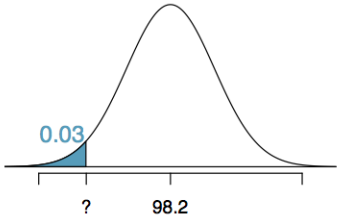
$$Z_{35.8} = \frac{35.8 - 36}{0.11} = -1.82$$

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

$$P(35.8 < X < 36.2) = P(-1.82 < Z < 1.82) = 0.9656 - 0.0$$

Example: 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

$.03 \longrightarrow P(Z < -1.88) = 0.03$

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

$$x = (-1.88 \times 0.73) + 98.2 = 96.8\,^\circ\text{F}$$

Practice

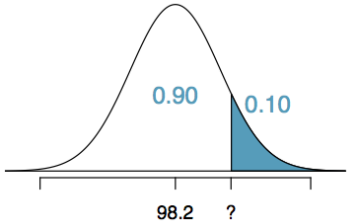
Body temperatures of healthy humans are distributed nearly normally with mean 98.2 ° F and standard deviation 0.73 ° F. What is the cutoff for the highest 10% of human body temperatures?

1. 97.3 ° F
2. 99.1 ° F
3. 99.4 ° F
4. 99.6 ° F

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?

1. 97.3 ° F
2. *99.1 ° F*
3. 99.4 ° F
4. 99.6 ° F



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
1.3	0.9115	0.9131	0.9147	0.9162	0.9177

$,10 \longrightarrow P(Z > 1.28) = 0.10$

$\text{bs} - \text{mean} \longrightarrow \frac{x - 98.2}{0.73} = 1.28$

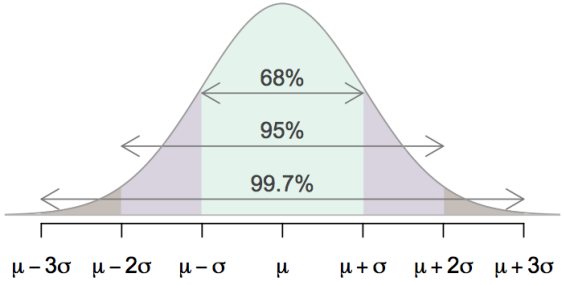
$.28 \times 0.73) + 98.2 = 99.1\ ^\circ\text{F}$

68-95-99.7 Rule

For 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

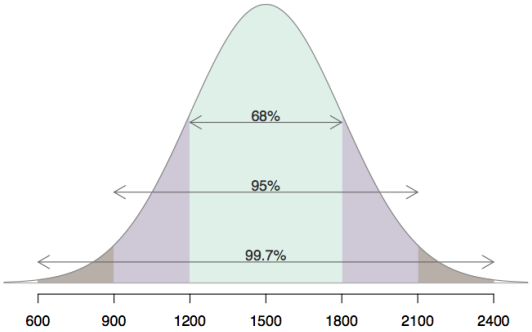
It is possible for observations to fall 4, 5 or even more standard deviations away from the mean, but these occurrences are very rare if the data are nearly normal.



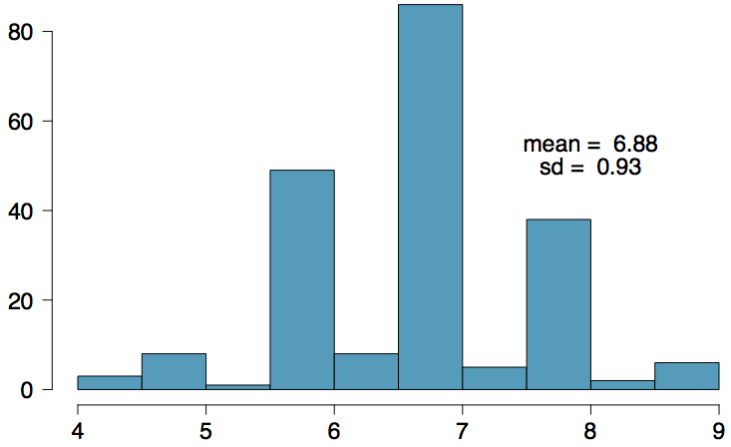
Describing variability using the 68-95-99.7 Rule

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

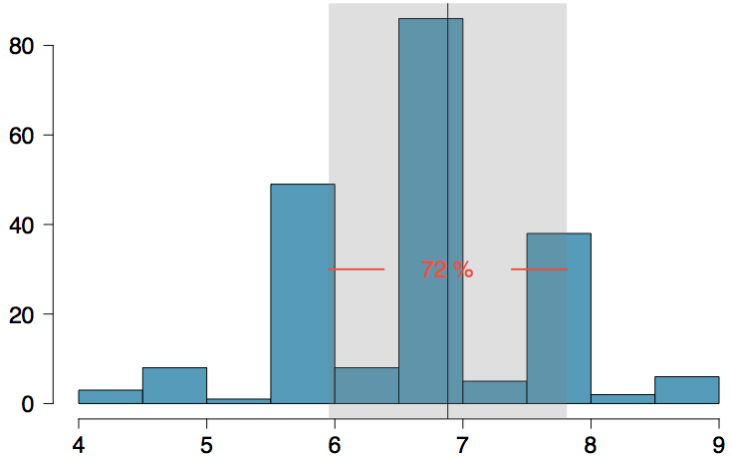
- ≈ 68% of students score between 1200 and 1800 on the SAT
- ≈ 95% of students score between 900 and 2100 on the SAT
- ≈ 99.7% of students score between 600 and 2400 on the SAT



Example: Number of nights of sleep on school nights

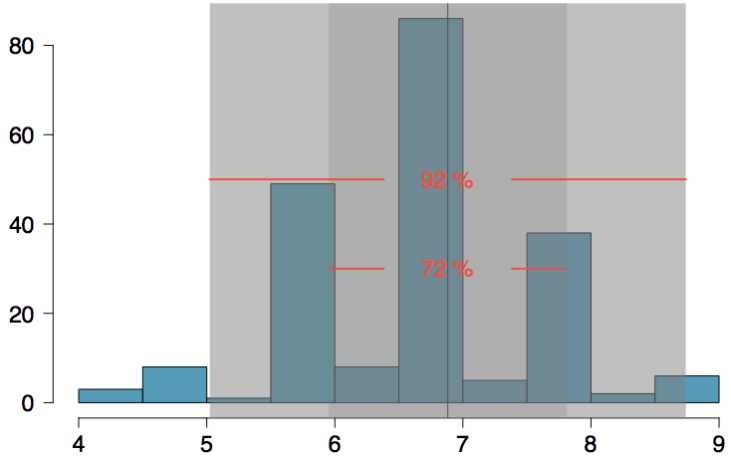


Example: Number of nights of sleep on school nights



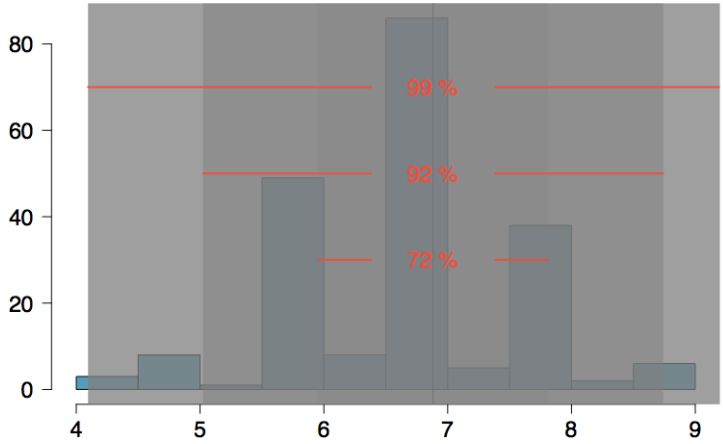
Mean = 6.88 hours, SD = 0.92 hours.
72% of the data are within 1 SD of the mean: 6.88 $\pm$ 0.93.

Example: Number of nights of sleep on school nights



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

Example: Number of nights of sleep on school nights



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

Practice

Which of the following is **false**?

1. Majority of Z scores in a right skewed distribution are negative.
2. In a skewed distributions the Z score of the mean might be different than 0.
3. For a normal distribution, IQR is less than 2 x SD.
4. Z scores are helpful for determining how unusual a data point is compared to the rest of the data in the distribution.

Practice

Which of the following is **false**?

1. Majority of Z scores in a right skewed distribution are negative.
2. In a skewed distributions the Z score of the mean might be different than 0.
3. *For a normal distribution, the IQR is less than 2 x SD.*
4. Z scores are helpful for determining how unusual a data point is compared to the rest of the data in the distribution.

The Normal Approximation

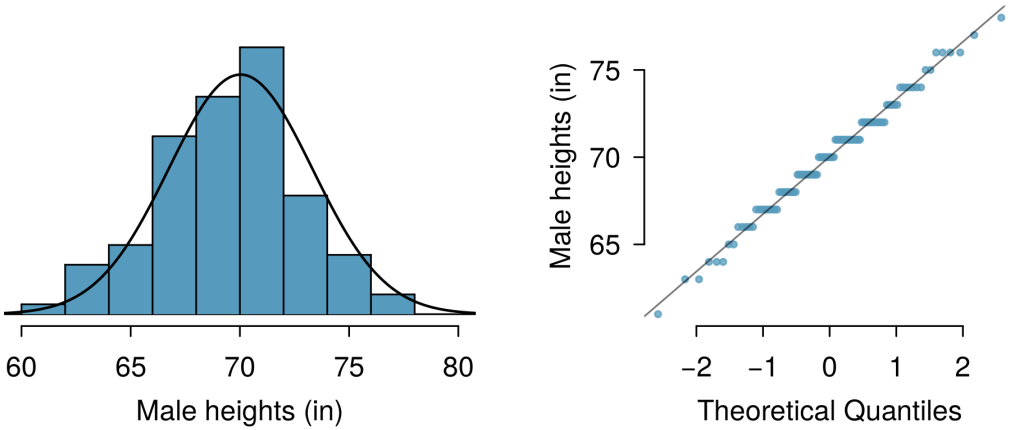
Evaluating The Normal Approximation

We often use the normal distribution as an **approximation**, taking real data and assuming it follows the normal.

How do we tell whether this is a good assumption?

Normal probability plot (QQ)

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



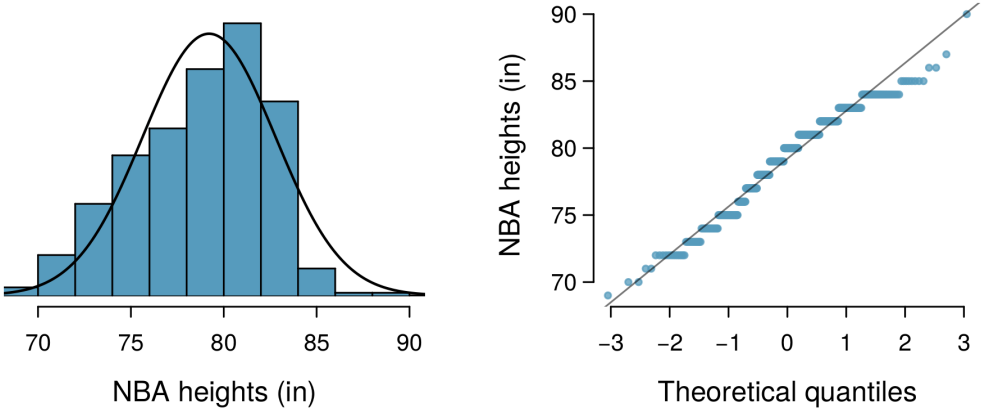
The right hand plot is also called a **quantile-quantile plot**, or **QQ plot** for short.

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 R when making these plots.

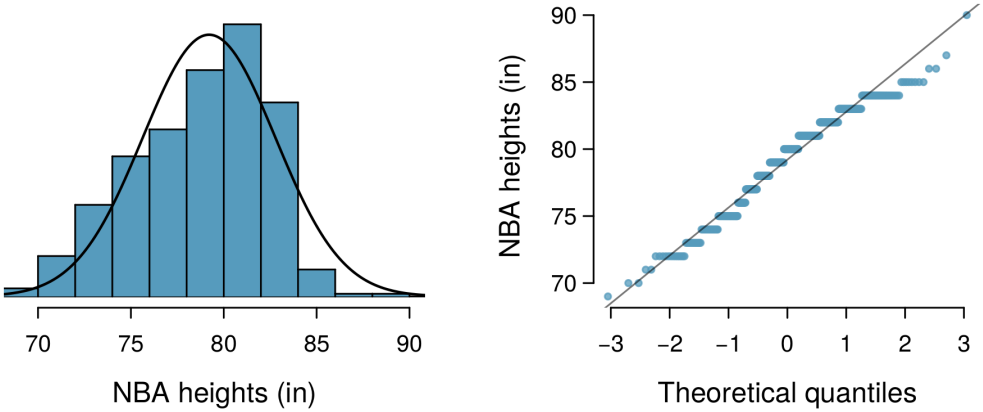
NBA Heights, 2008-09

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?



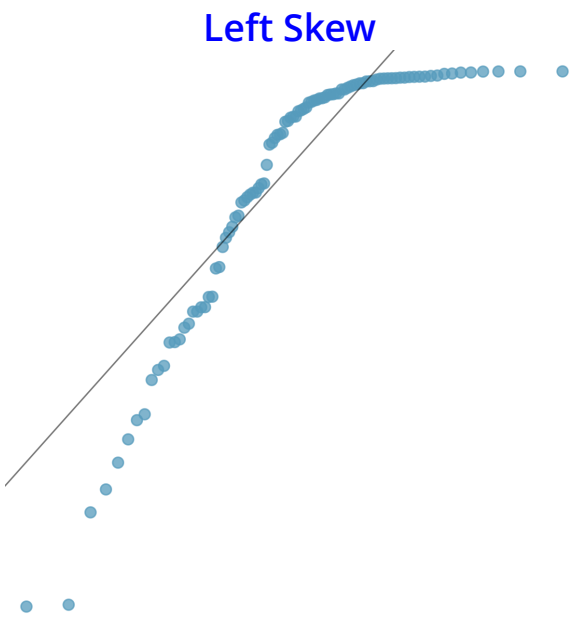
NBA Heights, 2008-09

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?



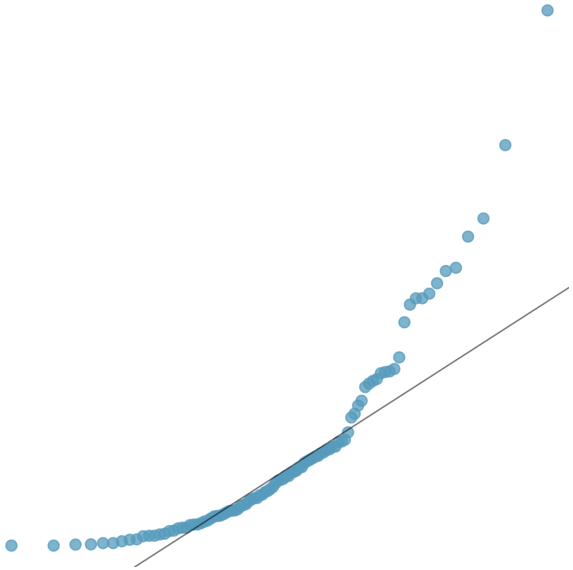
Why do the points on the normal probability have jumps?

Normal probability plot and skewness



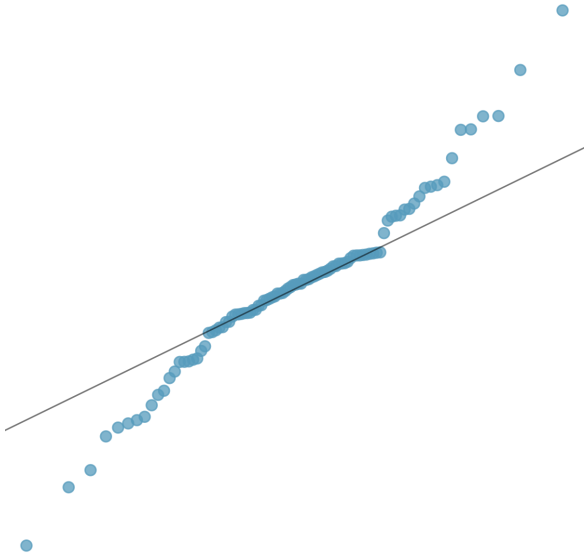
Normal probability plot and skewness

Right Skew



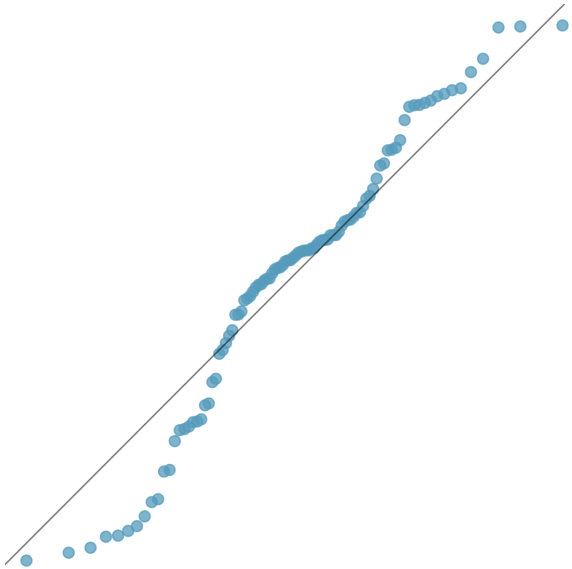
Normal probability plot and skewness

Long Tails



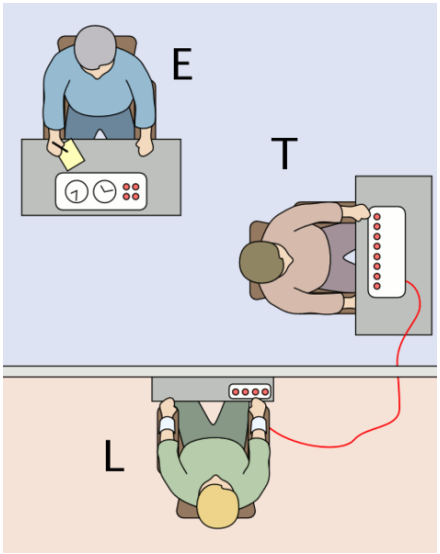
Normal probability plot and skewness

Short Tails



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.

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?

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

Geometric distribution

… the third person?

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

Geometric distribution

… the tenth person?

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

Geometric distribution (cont.)

The **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 distribution (cont.)

Geometric probabilities: If p represents probability of success, $(1 - p)$ represents probability of failure, and n represents number of independent trials
$P(\text{success on the } n^{th} \text{ trial}) = (1 - p)^{n-1}p.$

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.}

- no, on the roll of a die there are more than 2 possible outcomes
- yes, why not

Geometric distribution (cont.)

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.}

- no, on the roll of a die there are more than 2 possible outcomes
- *yes, why not*

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

Expected value

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

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 $\frac{1}{p}$.

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

Thus, we would expect her 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:

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

Expected value and its variability

- 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

Binomial: Redux

We've already talked about the binomial once, but let's revisit it a little more today, with some more details.

The Binomial Distribution

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

The Binomial Distribution

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":

- Scenario 1: A (refuse) $\times$ B (shock) $\times$ C (shock) $\times$ D (shock) = $0.35 \times 0.65^3 = 0.096$.
- Scenario 2: A (shock) $\times$ B (refuse) $\times$ C (shock) $\times$ D (shock) = $0.65 \times 0.35 \times 0.65^2 = 0.096$.
- Scenario 3: A (shock) $\times$ B (shock) $\times$ C (refuse) $\times$ D (shock) = $0.65^2 \times 0.35 \times 0.65 = 0.096$.
- Scenario 4: A (shock) $\times$ B (shock) $\times$ C (shock) $\times$ D (refuse) = $0.65^3 \times 0.35 = 0.096$.

The Binomial Distribution

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":

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

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

The 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

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

The 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

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

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

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

The Binomial distribution

So 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 $k = 2$:

RRSSSSSSS
S**RR**SSSSSS
…
SSSSSS**RR**S
SSSSSSS**RR**
…

then writing out all possible scenarios would be incredibly tedious and prone to errors. Remember: the **RR** don't have to be together. We have to figure out all the **R**SS**R** too, etc., etc.

Calculating the # of scenarios

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

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

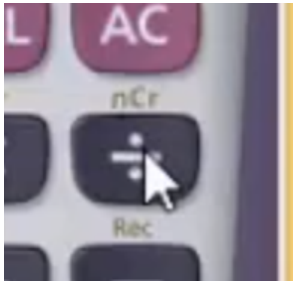
- $k = 1, n = 4$: $\binom{4}{1} = \frac{4!}{1!\,(4-1)!} = \frac{4\times3\times2\times1}{1\times(3\times2\times1)} = 4$
- $k = 2, n = 9$: $\binom{9}{2} = \frac{9!}{2!\,(9-1)!} = \frac{9\times8\times7!}{2\times1\times7!} = \frac{72}{2} = 36$

```
choose(9, 2)
```

```
## [1] 36
```

Calculating the # of scenarios

You can also do this on all scientific calculators, using the **nCr** button (or equivalent).



Properties of the choose function

Which of the following is false?

- There are n ways of getting 1 success in n trials, $\binom{n}{1} = n$.
- There is only 1 way of getting n successes in n trials, $\binom{n}{n} = 1$.
- There is only 1 way of getting n failures in n trials, $\binom{n}{0} = 1$.
- There are $n - 1$ ways of getting $n - 1$ successes in n trials, $\binom{n}{n-1} = n - 1$.

Properties of the choose function

Which of the following is false?

- There are n ways of getting 1 success in n trials, $\binom{n}{1} = n$.
- There is only 1 way of getting n successes in n trials, $\binom{n}{n} = 1$.
- There is only 1 way of getting n failures in n trials, $\binom{n}{0} = 1$.
- *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

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

Practice

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

1. the trials must be independent
2. the number of trials, n, must be fixed
3. each trial outcome must be classified as a *success* or a *failure*
4. the number of desired successes, k, must be greater than the number of trials
5. the probability of success, p, must be the same for each trial

Practice

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

1. the trials must be independent
2. the number of trials, n, must be fixed
3. each trial outcome must be classified as a *success* or a *failure*
4. **the number of desired successes, k, must be greater than the number of trials**
5. the probability of success, p, must be the same for each trial

Gallup Poll: Obesity in America

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?

- pretty high?
- pretty low?

Gallup Poll: Obesity in America

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?

- $0.262^8 \times 0.738^2$
- $\binom{8}{10} \times 0.262^8 \times 0.738^2$
- $\binom{10}{8} \times 0.262^8 \times 0.738^2$
- $\binom{10}{8} \times 0.262^2 \times 0.738^8$

Gallup Poll: Obesity in America

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?

- $0.262^8 \times 0.738^2$
- $\binom{8}{10} \times 0.262^8 \times 0.738^2$
- **$\binom{10}{8} \times 0.262^8 \times 0.738^2 = 45 \times 0.262^8 \times 0.738^2 = 0.0005$**
- $\binom{10}{8} \times 0.262^2 \times 0.738^8$

The birthday problem

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

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

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.

$$P(\text{no matches}) = 1 \times \left(1 - \frac{1}{365}\right) \times \left(1 - \frac{2}{365}\right) \times \cdots \times \left(1 - \frac{120}{365}\right)$$

The birthday problem (cont.)

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

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

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?

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.

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

Gallup Poll: Home Schooling

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?}

- Yes
- No

	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

Gallup Poll: Home Schooling

$$\mu = np = 1,000 \times 0.13 = 130$$
$$\sigma = \sqrt{np(1-p)} = \sqrt{1,000 \times 0.13 \times 0.87} \approx 10.6$$

- Method 1: Range of usual observations: $130 \pm 2 \times 10.6 = (108.8, 151.2)$
  (100 is outside this range, so would be considered unusual)
- Method 2: Z-score of observation: $Z = \frac{x - mean}{SD} = \frac{100 - 130}{10.6} = -2.83$
  (100 is more than 2 SD below the mean, so would be considered unusual)

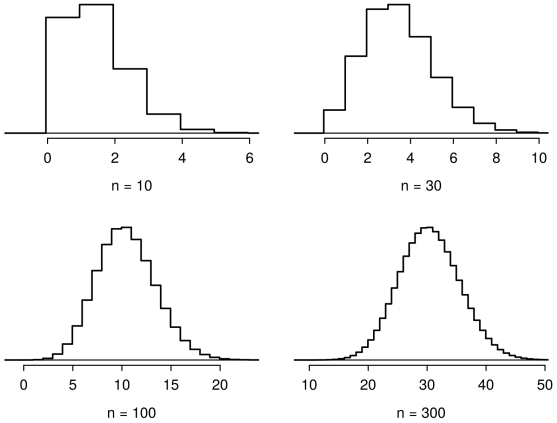
Normal approximation to the binomial (Practice)

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.

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

Practice

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

- $n = 100, p = 0.95$
- $n = 25, p = 0.45$
- $n = 150, p = 0.05$
- $n = 500, p = 0.015$

Practice

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

- $n = 100, p = 0.95$
- **$n = 25, p = 0.45 \rightarrow 25 \times 0.45 = 11.25; 25 \times 0.55 = 13.75$**
- $n = 150, p = 0.05$
- $n = 500, p = 0.015$

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?

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?

Power users contribute much more content than the typical user.

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.

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.

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

This seems like an awful lot of work…

Could Use R ...

```
sum(dbinom(x = 70:245, size = 245, prob = 0.25))
```

```
## [1] 0.112763
```

```
pbinom(q = 69, size = 245, prob = 0.25, lower.tail = FALSE)
```

```
## [1] 0.112763
```

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 $p = 0.25$.

$$\mu = 245 \times 0.25 = 61.25 \qquad \sigma = \sqrt{245 \times 0.25 \times 0.75} = 6.778$$

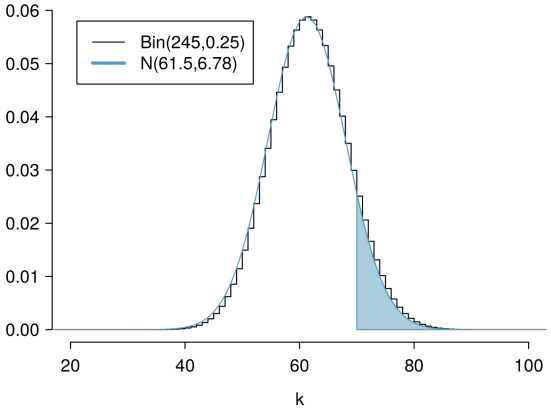
- $Bin(n = 245, p = 0.25) \approx N(\mu = 61.25, \sigma = 6.778)$.

Normal approximation to the binomial

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

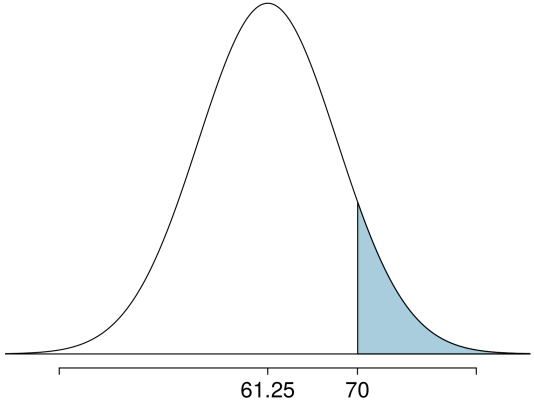
$$\mu = 245 \times 0.25 = 61.25 \qquad \sigma = \sqrt{245 \times 0.25 \times 0.75} = 6.778$$

- $Bin(n = 245, p = 0.25) \approx N(\mu = 61.25, \sigma = 6.778)$.



Computing the Approximation

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



$$Z = \frac{obs - mean}{SD} = \frac{70 - 61.25}{6.778}$$
$$= 1.29094$$
$$P(Z > 1.29094) = 1 - 0.90164$$
$$= 0.09836$$

Computing the Approximation

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

But where did this $P(Z > 1.29)$ answer come from? R again!

$$Z = \frac{obs - mean}{SD} = \frac{70 - 61.25}{6.778}$$
$$= 1.29094$$
$$P(Z > 1.29094) = 1 - 0.90164$$
$$= 0.09836$$

Computing Normal Probabilities

Just like we did for **pbinom()** and **dbinom()**, we can do for **pnorm()** and **dnorm()**. You saw this in workshop last week.

```
pnorm(1.290964, lower.tail = FALSE)
```

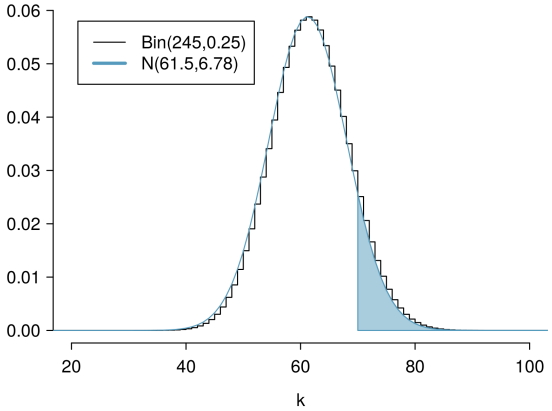
```
## [1] 0.09835808
```

This seems … bad

We know the exact probability, done using **pbinom()**, is 0.1128. So why is this "approximation" giving an answer of 0.09836?

This seems … bad

We know the exact probability, done using **pbinom()**, is 0.1128. So why is this "approximation" giving an answer of 0.09836?



“Correction for Continuity”

The normal approximation to the binomial can be a little rough. There is a **correction for continuity** which can be used instead:

- The cutoff values for the lower end of a shaded region should be reduced by 0.5
- The cutoff values for the upper end should be increased by 0.5.

Since we are doing a “greater than” probability, the lower end of the shaded region is our relevant object, so we reduce.

Computing Normal Probabilities

$$
\begin{aligned}
Z &= \frac{obs - mean}{SD} = \frac{(70 - 0.5) - 61.25}{6.778} \\
&= 1.217173 \\
P(Z > 1.217173) &= 1 - 0.8882308 \\
&= 0.1117692
\end{aligned}
$$

```
pnorm(1.217173, lower.tail = FALSE)
```

```
## [1] 0.1117692
```

That's much better!

Inverse Normal Problems

The Inverse Problem

As we've seen in the previous, we often need to take Z-scores and find probabilities from them. Sometimes this is in a normal problem, sometimes an approximation, and so on.

What if you **wanted to go backward**? What would this look like?

Inverse Problem Statement

What if I told you “the probability of this event happening is 0.5”. What would the Z-score of such a setup be?

$$P(Z \leq z_0) = 0.5$$

What’s the unknown here? $\boldsymbol{z_0}$!

How do we solve for z_0?

- trial and error?
- use tables?
- **use R!**

Using R to Find z_0

```
qnorm(p = 0.5, mean = 0, sd = 1, lower.tail = TRUE)
```

```
## [1] 0
```

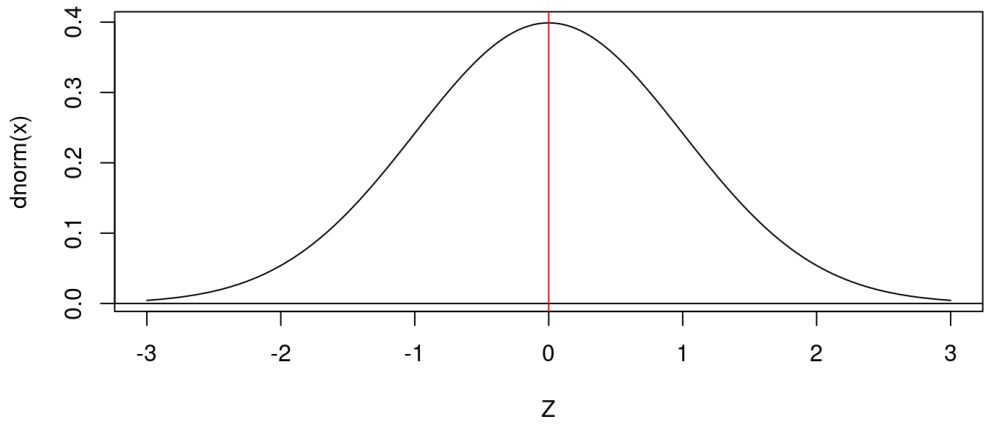
Does this make sense? This is saying that $z_0 = 0$ has probability to its left of 0.5.

```
pnorm(q = 0, mean = 0, sd = 1, lower.tail = TRUE)
```

```
## [1] 0.5
```

Checking Again

```
x <- seq(from = -3, to = 3, by = 0.01)
plot(x, dnorm(x), type = "l", xlab = "Z")
abline(h = 0)
abline(v = 0, col = "red")
```

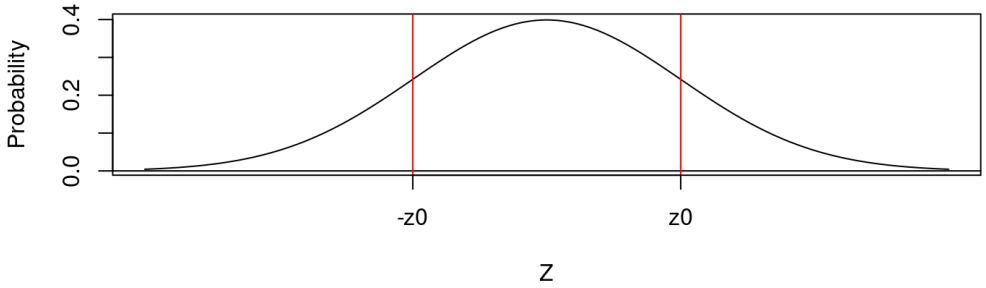


Practice

What value of z_0 has $P(-z_0 \leq Z \leq z_0) = 0.3$?

Practice

What value of z_0 has $P(-z_0 \leq Z \leq z_0) = 0.3$?



So what do we actually need to run?

Practice

```
qnorm(p = 0.15, mean = 0, sd = 1, lower.tail = TRUE)
```

```
## [1] -1.036433
```

This is **-z0**, because we used 0.15 area to the left, and **lower.tail = TRUE**.

```
qnorm(p = 0.15, mean = 0, sd = 1, lower.tail = FALSE)
```

```
## [1] 1.036433
```

And this is z_0, because **lower.tail = FALSE**.