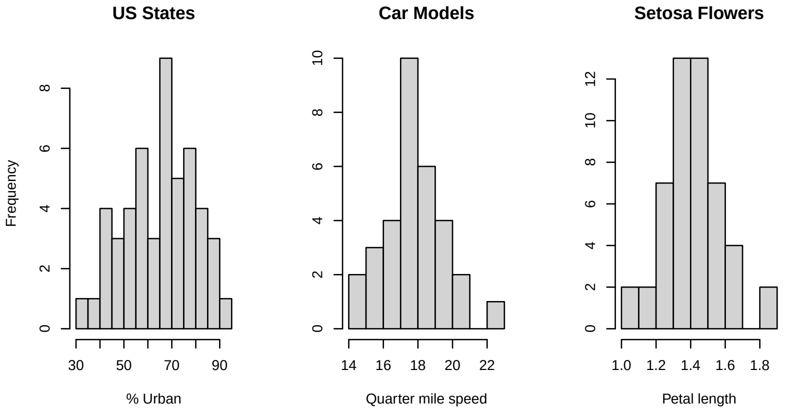
Normal Approximations

Ryan Miller

Different but Similar Data?



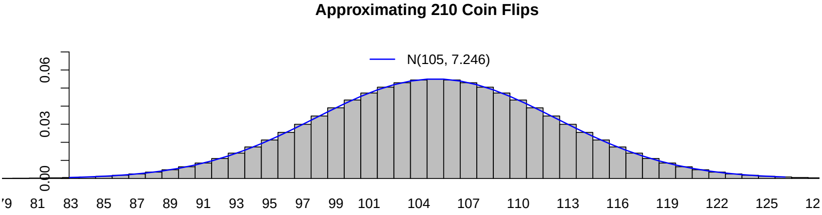
Practically speaking, these data are very different, but do you see any similarities (ignoring the units)?

The Normal Distribution

- Last time, we introduced the idea of approximating another distribution with a normal curve
  - To make this work, we needed proper values of μ and σ, the parameters dictating the normal curve's *center* and *spread*

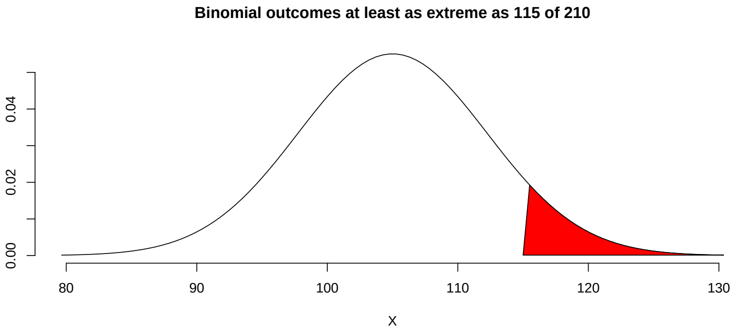
The Normal Distribution

- Last time, we introduced the idea of approximating another distribution with a normal curve
  - To make this work, we needed proper values of μ and σ, the parameters dictating the normal curve's *center* and *spread*
- We used the random variable's *expected value* and *standard deviation* (square root of variance)
  - For a *binomial random variable*, $E(X) = n * p$ and $StdDev(X) = \sqrt{n * p * (1 - p)}$



Probability and the Normal Curve

- The normal distribution is useful because the area under it can be used to find probabilities
  - Shaded in red is the area representing $P(X \geq 115)$
  - This area *approximates* the one-sided p-value for 115 successes in 210 trials under the null model that $p = 0.5$



Probability and the Normal Curve

- Recall we could have calculated this p-value *exactly* using the *binomial distribution* by summing many different discrete probabilities, $P(X = 115) + P(X = 116) + \ldots + P(X = 210)$
  - But would this summation approach work if X were continuous?

Probability and the Normal Curve

- Recall we could have calculated this *p*-value *exactly* using the *binomial distribution* by summing many different discrete probabilities, $P(X = 115) + P(X = 116) + \ldots + P(X = 210)$
  - But would this summation approach work if X were continuous?
- For continuous random variables, it *only* makes sense calculate probabilities using areas

Probability and the Normal Curve

- Recall we could have calculated this *p*-value *exactly* using the *binomial distribution* by summing many different discrete probabilities, $P(X = 115) + P(X = 116) + \ldots + P(X = 210)$
  - But would this summation approach work if X were continuous?
- For continuous random variables, it *only* makes sense calculate probabilities using areas
- A continuous variable can theoretically be measured with infinite precision
  - Thus, the probability of observing an average height of precisely 70.25 inches in a sample is zero (because there are infinitely many possibilities!)
  - However, the probability of observing an average height in the interval $70.25 \pm \epsilon$ is calculable

Functions in R

- `pnorm()` is the primary `R` function for *calculating probabilities* using the normal distribution, the main arguments are:
  - `q` - the boundary value defining your area (ie: $x = 115$)
  - `mean` - the normal distribution's mean (ie: $\mu = 105$)
  - `sd` - the distributions standard deviation (ie: $\sigma = 7.246$)
  - `lower.tail` - whether to take the area to the left of the boundary value (TRUE) or to the right of the boundary value (FALSE)

```
pnorm(115, mean = 105, sd = 7.246, lower.tail = TRUE)
```

```
## [1] 0.9162177
```

```
pnorm(115, mean = 105, sd = 7.246, lower.tail = FALSE)
```

```
## [1] 0.08378228
```

```
pnorm(115, mean = 115, sd = 7.246,lower.tail = FALSE)
```

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

Example

- The National Health and Nutrition Examination Survey (NHANES) is a national study designed assess US population health and nutrition
- The NHANES sample included 2649 adult women
  - The average height was 63.5 inches
  - The standard deviation of these heights was 2.75 inches

Question: Suppose you have a friend who is very shallow, and is only interested in dating adult women who are between 5'3 (63 in) and 5'6 (66 in). First, describe this scenario using a random variable, X. Then, use a normal approximation of the NHANES data to estimate the probability that a randomly selected adult female is between 5'3 and 5'6.

Example (solution)

- Let the random variable X denote the height of a randomly chosen adult female
  - We want to determine $P(63 \leq X \leq 66)$
- R can provide us with the areas to the left of 63 and 66 inches, then area in between these values can be found via subtraction:

```
p66 <- pnorm(66, mean = 63.5, sd = 2.75, lower.tail = TRUE)
p63 <- pnorm(63, mean = 63.5, sd = 2.75, lower.tail = TRUE)
p66 - p63
```

```
## [1] 0.3904862
```

- So, we estimate there's a 39.0% chance that a randomly selected women meets our friend's height criteria
  - In case you're curious, in the actual NHANES sample, 38.8% of women were in this range

Standardization

- While we could calculate probabilities using any normal curve, doing so (without software like `R`) is non-trivial because the normal curve does not have a closed-form integral

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

Standardization

- While we could calculate probabilities using any normal curve, doing so (without software like `R`) is non-trivial because the normal curve does not have a closed-form integral

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

- For mostly historical reasons, this has led statisticians to **standardize** their data in order to make use of the $N(0,1)$ curve
  - Until recently, statistics textbooks devoted many pages to tables detailing various areas under $N(0,1)$ curve
  - Modern statistical software has made using such tables obsolete

Normal Table

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Source: Abridged from Table 1 of A. Hald, *Statistical Tables and Formulas* (New York: John Wiley & Sons, Inc.), 1952. Reproduced by permission of the publisher.

Z-Scores

Suppose a random variable follows a normal distribution (ie: $X \sim N(\mu, \sigma)$), then:

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

- These transformed, or *standardized*, values are called Z-scores
  - A Z-score can be interpreted as how many standard deviations an observed data-point is above or below its expected value

Z-Scores

Suppose a random variable follows a normal distribution (ie: $X \sim N(\mu, \sigma)$), then:

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

- These transformed, or *standardized*, values are called Z-scores
  - A Z-score can be interpreted as how many standard deviations an observed data-point is above or below its expected value
- For example, suppose X is a random variable from a $N(\mu = 3.13, \sigma = 2.23)$ distribution and we observe $x = 5.19$

Z-Scores

Suppose a random variable follows a normal distribution (ie: $X \sim N(\mu, \sigma)$), then:

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

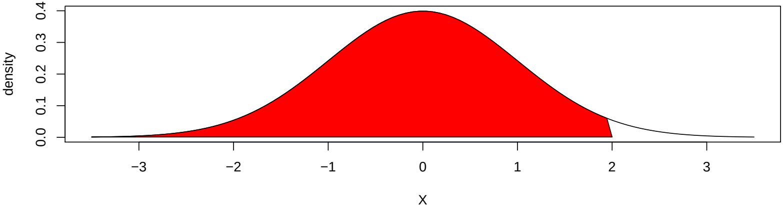
- These transformed, or *standardized*, values are called Z-scores
  - A Z-score can be interpreted as how many standard deviations an observed data-point is above or below its expected value
- For example, suppose X is a random variable from a $N(\mu = 3.13, \sigma = 2.23)$ distribution and we observe $x = 5.19$
  - The corresponding Z-score is $z = (5.19 - 3.13)/2.23 = 0.923$
  - So this data-point is slightly less than one standard deviation above average (meaning it is quite typical)

Z-scores (interpretation)

- Z-scores can be useful in helping non-experts understand variables with obtuse units
  - For example, if a doctor tells me that my blood urea nitrogen is 8 nmol/L above average I don't know if I should worry
  - But if they tell me it's 4 standard deviations above average I know I should be concerned

Z-scores (interpretation)

- Z-scores can be useful in helping non-experts understand variables with obtuse units
  - For example, if a doctor tells me that my blood urea nitrogen is 8 nmol/L above average I don't know if I should worry
  - But if they tell me it's 4 standard deviations above average I know I should be concerned
- Z-scores can also be mapped to *percentiles* of the normal distribution
  - A Z-score of 2 indicates a value larger than 97.7% of the data-points (we'll see how to find this in `R` next)



Functions in R

- `qnorm()` allows to *find percentiles* of a normal distribution its main arguments are:
  - `p` - the percentile of interest
  - `mean` - the normal distribution's mean (ie: $\mu = 0$)
  - `sd` - the distributions standard deviation (ie: $\sigma = 1$)

```
qnorm(.5, mean = 0, sd = 1)
```

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

```
qnorm(.75, mean = 0, sd = 1)
```

```
## [1] 0.6744898
```

```
qnorm(.99, mean = 0, sd = 1)
```

```
## [1] 2.326348
```

Example (revisited)

Recall the NHANES sample included 2649 adult women, who had an average height was 63.5 inches with a standard deviation 2.75 inches

1) For our very shallow friend who was interested in dating adult women who are between 5'3 (63 in) and 5'6 (66 in), use Z-scores and the *standard normal distribution* to express and calculate the probability of a randomly chosen women falling within this range.
2) Suppose a different (but equally shallow) friend wants to maximize his chances of having a son who plays college basketball, thus he is only interested in dating women who are above the 95% percentile for height. Use the NHANES sample and a normal approximation to determine the 95^{th} percentile.

Example (solution #1)

- Let X be a random variable describing a randomly chosen female's height.
  - The Z-score for our friend's lower threshold is $z_L = \frac{63-63.5}{2.75} = -0.18$
  - The upper threshold is $z_U = \frac{66-63.5}{2.75} = 0.91$
- Thus, $P(63 \leq X \leq 66) = P(-0.18 \leq Z \leq 0.91)$, which we can find using `R` using `pnorm()`:

```
pzu <- pnorm(.91, mean = 0, sd = 1, lower.tail = TRUE)
pzl <- pnorm(-.18, mean = 0, sd = 1, lower.tail = TRUE)
pzu - pzl
```

```
## [1] 0.3900125
```

Example (solution #2)

We could can find the Z-score corresponding to the 95% percentile using R using `qnorm()`, then work backwards (un-standardize) to find x

```
qnorm(.95, mean = 0, sd = 1)
```

```
## [1] 1.644854
```

$$1.645 = \frac{x - 63.5}{2.75} \implies x = 1.645 * 2.75 + 63.5 = 68.02$$

Alternatively, we could specify the mean/sd of our data in `qnorm()`

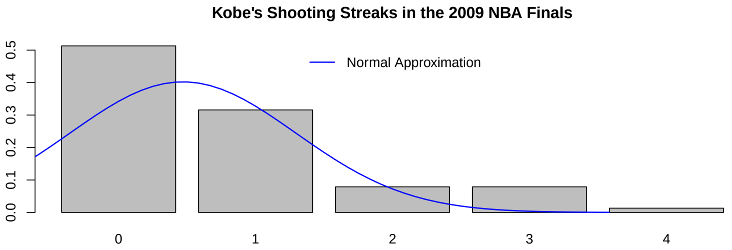
```
qnorm(.95, mean = 63.5, sd = 2.75)
```

```
## [1] 68.02335
```

Either way, we see the 95^{th} percentile is roughly 5‘8

Approximation Accuracy

- In the two examples we've seen so far, approximating via the normal distribution has been remarkably accurate
- Unfortunately, the normal curve cannot be applied to all situations
  - The graph below shows a normal approximation of Kobe's shooting streaks, would you feel comfortable using this curve to estimate the probability of 3+ shot shooting streak?



Conclusion

- It turns out, there is something special going on in our two examples (binomial random variable with $n = 210$ trials, and adult female heights) that led to the normal approximation being accurate

Conclusion

- It turns out, there is something special going on in our two examples (binomial random variable with $n = 210$ trials, and adult female heights) that led to the normal approximation being accurate
- Next time, we'll explore a landmark theoretical result that provides tremendous insight regarding scenarios where a normal distribution will provide an accurate approximation
  - In fact, statisticians have *proven* the normal distribution to apply to certain scenarios

Conclusion

- It turns out, there is something special going on in our two examples (binomial random variable with $n = 210$ trials, and adult female heights) that led to the normal approximation being accurate
- Next time, we'll explore a landmark theoretical result that provides tremendous insight regarding scenarios where a normal distribution will provide an accurate approximation
  - In fact, statisticians have *proven* the normal distribution to apply to certain scenarios
- From this presentation, you should take away three things:
  - Be able to use the normal distribution to calculate probabilities
  - Understand Z-scores, standardization, and the standard normal distribution
  - Be able to find values representing certain percentiles of a normally distributed random variable