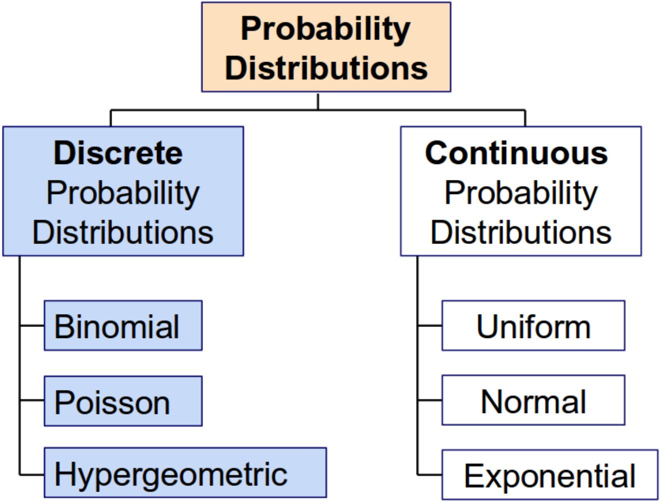
STATISTICS

Marco Caserta
marco.caserta@ie.edu

IE University

1. Continuous Distributions
2. Uniform Distribution
3. Normal distribution
4. Evaluating the normal approximation
5. Normal Approximation to the Binomial
6. Exponential Distribution
7. Joint Cumulative Distribution Functions

Probability Distributions


Probability Distributions
Discrete Probability Distributions
Continuous Probability Distributions
Binomial
Poisson
Hypergeometric
Uniform
Normal
Exponential

1. Continuous Distributions

2. Uniform Distribution

3. Normal distribution
   - Normal distribution model
   - Standardizing with Z scores
   - Normal probability table
   - Normal probability examples
   - 68-95-99.7 rule

4. Evaluating the normal approximation
   - Normal probability plot

5. Normal Approximation to the Binomial

6. Exponential Distribution

7. Joint Cumulative Distribution Functions

Continuous Probability Distribution Function

- A continuous random variable is a variable that can assume any value in an interval
- These can potentially take on any value, depending only on the ability to measure accurately

Cumulative Distribution Function

The cumulative distribution function, F(x), for a continuous random variable X expresses the probability that X does not exceed the value of x

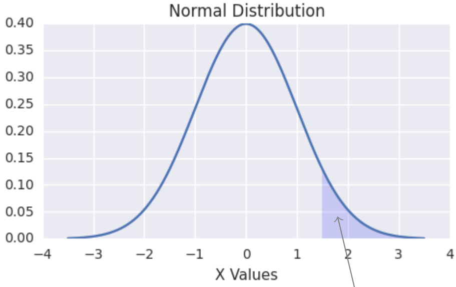
$$F(x) = p(X \leq x)$$

Let a and b be two possible values of X, with $a < b$. The probability that X lies between a and b is:

$$p(a < X < b) = F(b) - F(a)$$

Probability as an Area

- Shaded area under the curve is the probability that X is between a and b



$p(a \leq x \leq b) =$
$p(a < x < b)$

Probability Density Function

- The probability density function, $f(x)$, of random variable X has the following properties:
  - $f(x) > 0$ for all values of x
  - The area under the probability density function $f(x)$ over all values of the random variable X is equal to 1.0
  - The probability that X lies between two values is the area under the density function graph between the two values
  - The cumulative density function $F(x_0)$ is the area under the probability density function $f(x)$ from the minimum value x_m up to x_0:

$$\int_{x_m}^{x_0} f(x)dx$$

Mean and Variance of a Continuous Random Variable

- The mean is:

$$\mu_x = E(X) = \int_{-\infty}^{+\infty} x f(x) dx$$

- The variance is:

$$\sigma^2 = E[(X - \mu_x)^2]$$

- Remember the results for a linear function of variables $W = a + bY$:
  - $\mu_W = E(a + bY) = a + b\mu_x$
  - $\sigma^2 = Var(a + bX) = b^2\sigma_x^2$
- Apply the results to the special statistic:

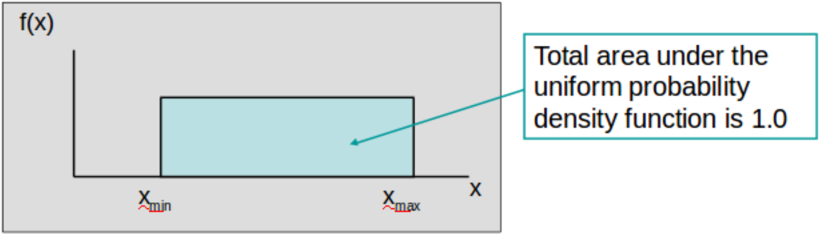
$$Z = \frac{X - \mu_x}{\sigma_x}$$

  - mean :
  - variance :

1. Continuous Distributions
2. Uniform Distribution
3. Normal distribution
   - Normal distribution model
   - Standardizing with Z scores
   - Normal probability table
   - Normal probability examples
   - 68-95-99.7 rule
4. Evaluating the normal approximation
   - Normal probability plot
5. Normal Approximation to the Binomial
6. Exponential Distribution
7. Joint Cumulative Distribution Functions

Uniform Distribution

- The uniform distribution is a probability distribution that has equal probabilities for all possible outcomes of the random variable



Total area under the uniform probability density function is 1.0

- The density function is:

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

- The mean is:

$$\mu = \frac{a+b}{2}$$

- The variance is:

$$\sigma^2 = \frac{(b-a)^2}{12}$$

1 Continuous Distributions

2 Uniform Distribution

3 Normal distribution

Normal distribution model
Standardizing with Z scores
Normal probability table
Normal probability examples
68-95-99.7 rule

4 Evaluating the normal approximation

Normal probability plot

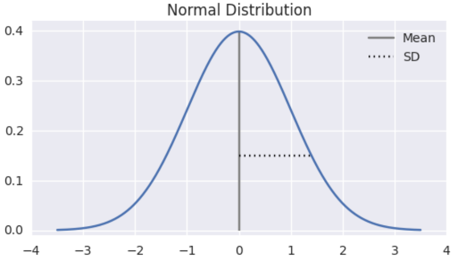
5 Normal Approximation to the Binomial

6 Exponential Distribution

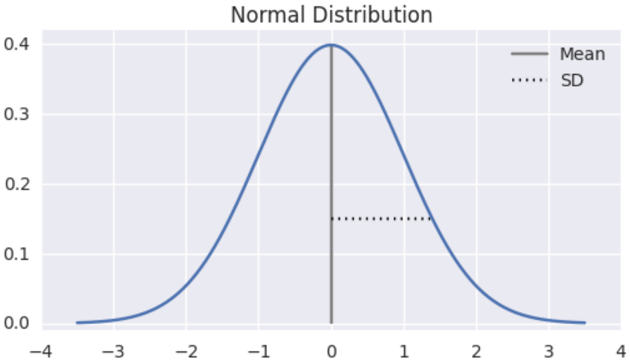
7 Joint Cumulative Distribution Functions

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



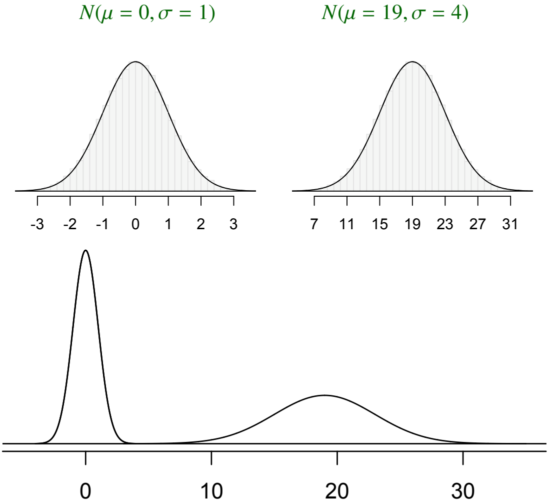
$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$



- Bell-shaped
- Mean = Median = Mode
- Symmetric
- Location is determined by the mean μ
- Spread (width) is determined by the standard deviation σ

Normal distributions with different parameters

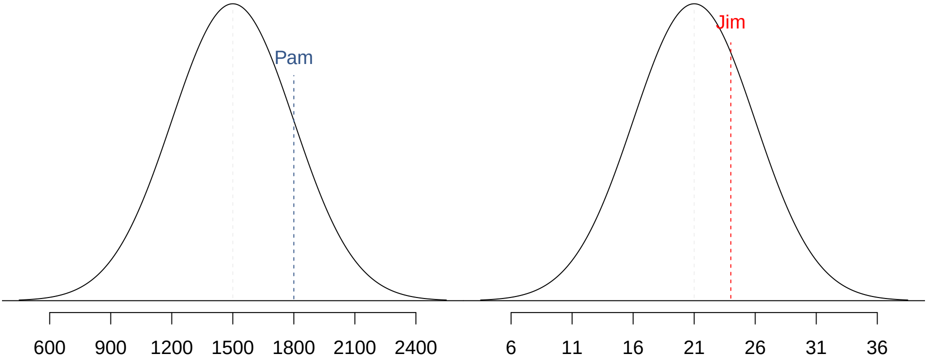
μ: mean, σ: standard deviation



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

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

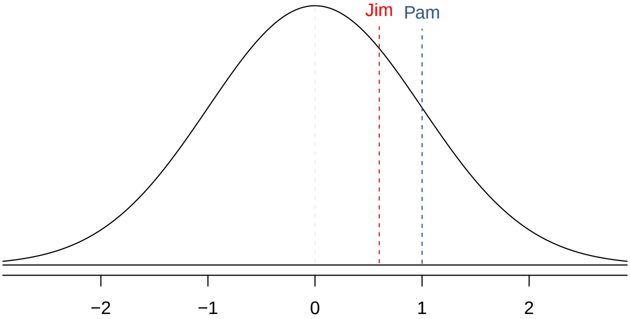
SAT scores are distributed nearly normally with mean 1500 and standard deviation 300. ACT scores are distributed nearly normally with mean 21 and standard deviation 5. A college admissions officer wants to determine which of the two applicants scored better on their standardized test with respect to 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.)

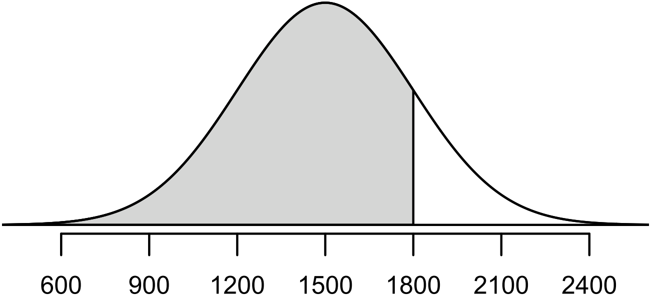
- These are called standardized scores, or Z scores.
- Z score of an observation is the number of standard deviations it falls above or below the mean.

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

- 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 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

$$p(x \leq 1.00) = 0.84$$

Excel → `NORMSDIST(1.00)` = 0.84

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?

Quality control

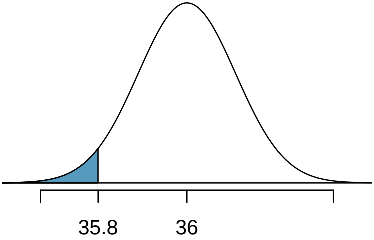
At Heinz ketchup factory the amounts which go into bottles of ketchup are supposed to be normally distributed with mean 36 oz. and standard deviation 0.11 oz. Once every 30 minutes a bottle is selected from the production line, and its contents are noted precisely. If the amount of ketchup in the bottle 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: $X \sim N(\mu = 36, \sigma = 0.11)$

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?

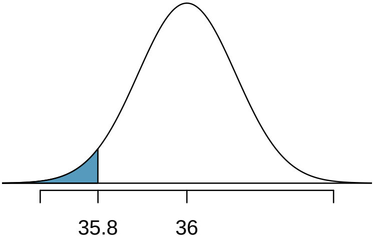
Let X = amount of ketchup in a bottle: $X \sim N(\mu = 36, \sigma = 0.11)$



Quality control

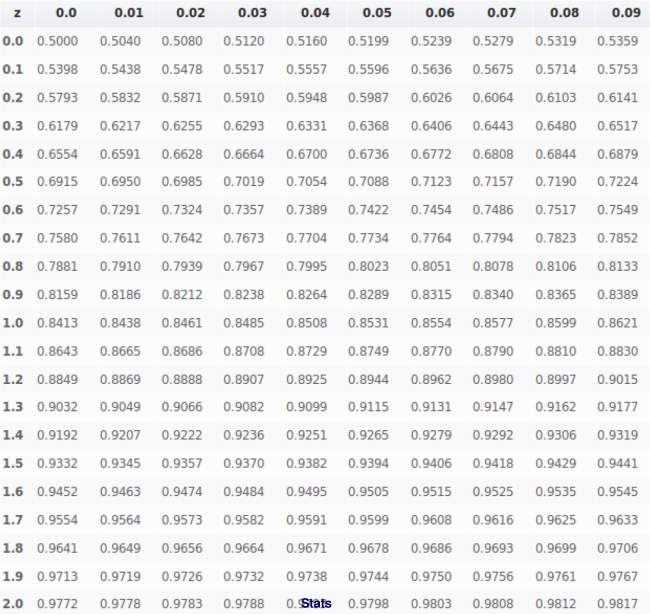
At Heinz ketchup factory the amounts which go into bottles of ketchup are supposed to be normally distributed with mean 36 oz. and standard deviation 0.11 oz. Once every 30 minutes a bottle is selected from the production line, and its contents are noted precisely. If the amount of ketchup in the bottle 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:* $X \sim N(\mu = 36, \sigma = 0.11)$



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

Finding the exact probability - using the Z table



z	0.0	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
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817

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%

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%

Practice

What percent of bottles pass the quality control inspection?

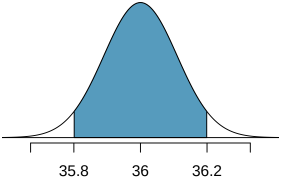
(a) 1.82%

(b) 3.44%

(c) 6.88%

(d) *93.12%*

(e) 96.56%



=

Practice

What percent of bottles pass the quality control inspection?

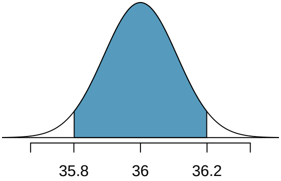
(a) 1.82%

(b) 3.44%

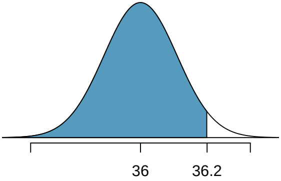
(c) 6.88%

(d) *93.12%*

(e) 96.56%



=



-

Practice

What percent of bottles pass the quality control inspection?

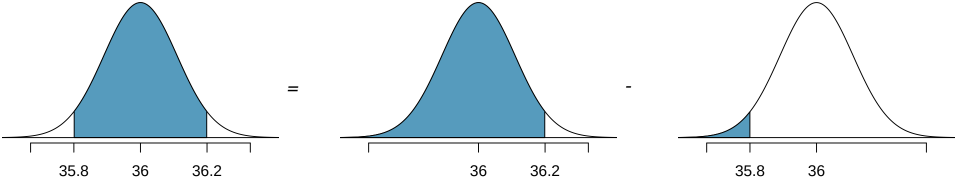
(A) 1.82%

(B) 3.44%

(C) 6.88%

(D) *93.12%*

(E) 96.56%



Practice

What percent of bottles pass the quality control inspection?

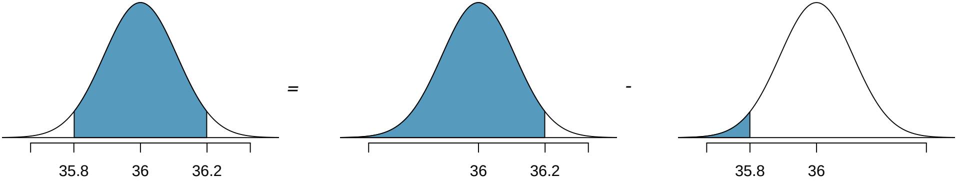
(a) 1.82%

(b) 3.44%

(c) 6.88%

(d) *93.12%*

(e) 96.56%



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

Practice

What percent of bottles pass the quality control inspection?

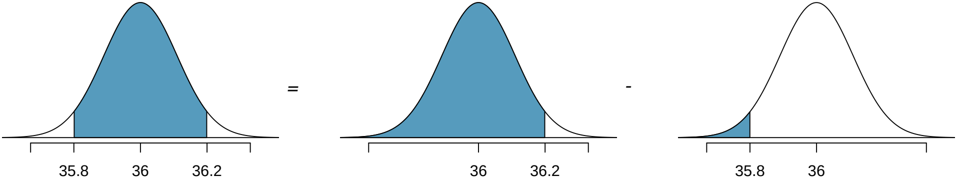
(a) 1.82%

(b) 3.44%

(c) 6.88%

(d) *93.12%*

(e) 96.56%



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

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

Practice

What percent of bottles pass the quality control inspection?

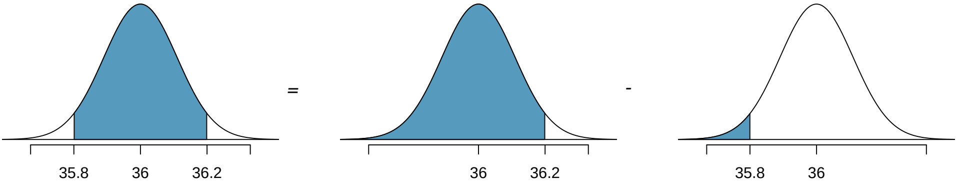
(A) 1.82%

(B) 3.44%

(C) 6.88%

(D) *93.12%*

(E) 96.56%



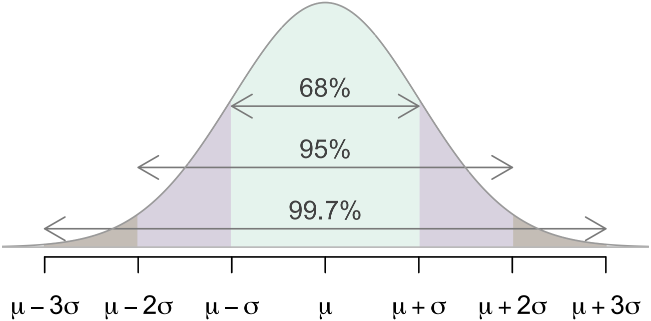
$$Z_{35.8} = \frac{35.8 - 36}{0.11} = -1.82$$

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

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

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.



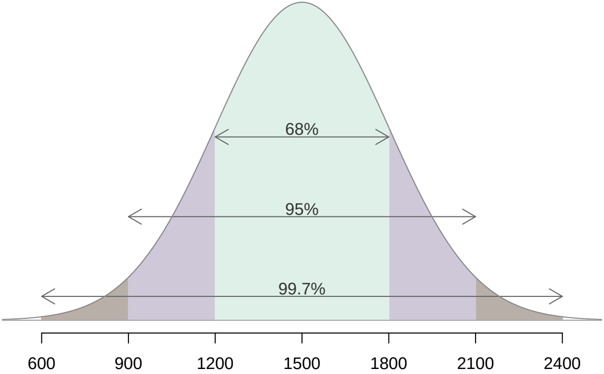
Describing variability using the 68-95-99.7 Rule

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

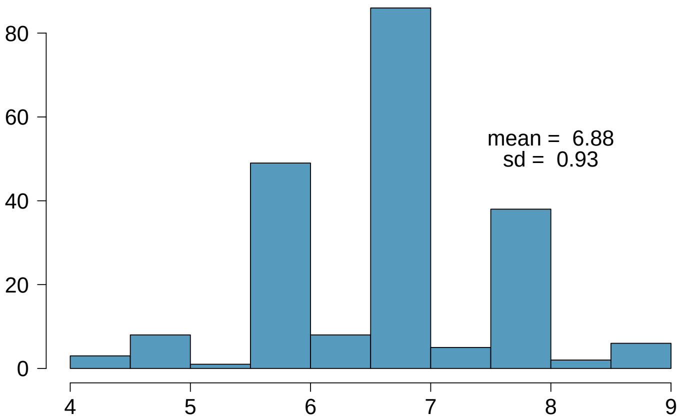
Describing variability using the 68-95-99.7 Rule

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.



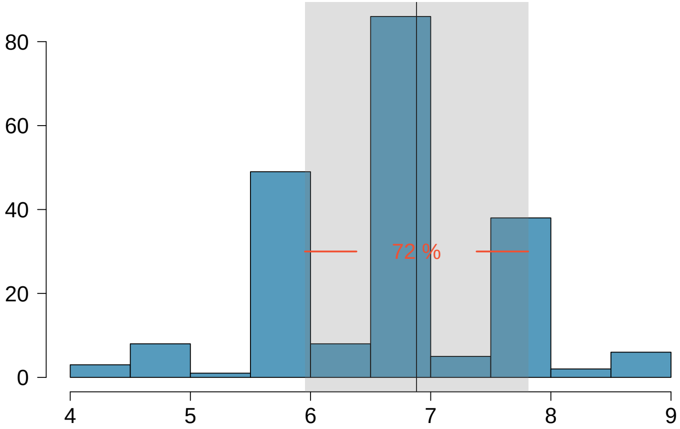
Number of hours of sleep on school nights



- Mean = 6.88 hours, SD = 0.92 hrs

$$6.88 \pm 0.93$$
$$6.88 \pm 2 \times 0.93$$
$$6.88 \pm 3 \times 0.93$$

Number of hours of sleep on school nights

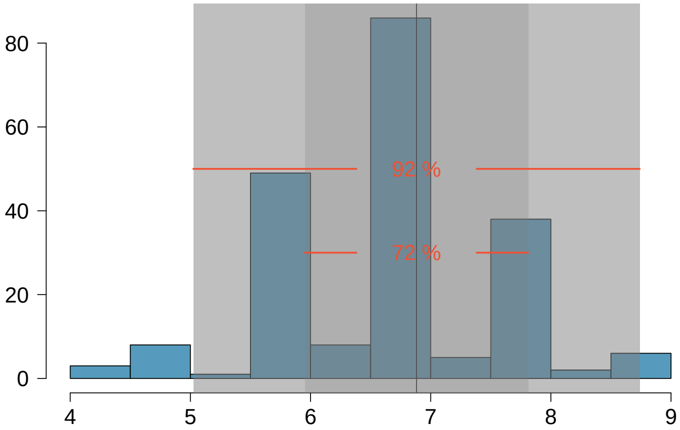


- Mean = 6.88 hours, SD = 0.92 hrs
- 72% of the data are within 1 SD of the mean: 6.88 ± 0.93

$6.88 \pm 2 \times 0.93$

$6.88 \pm 3 \times 0.93$

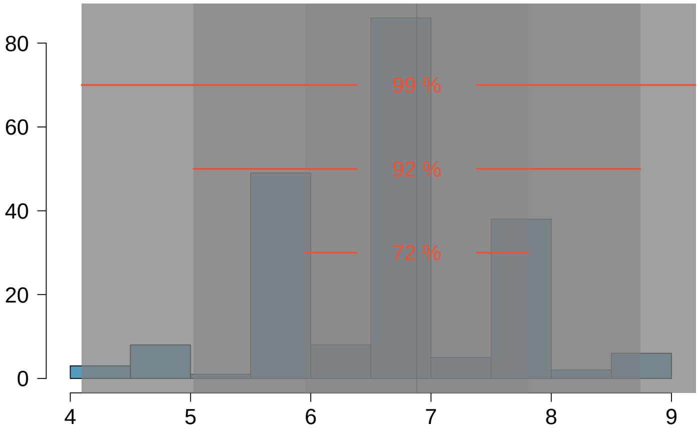
Number of hours of sleep on school nights



- 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 2 SD of the mean: $6.88 \pm 2 \times 0.93$

$6.88 \pm 3 \times 0.93$

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 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?

(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.

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.*

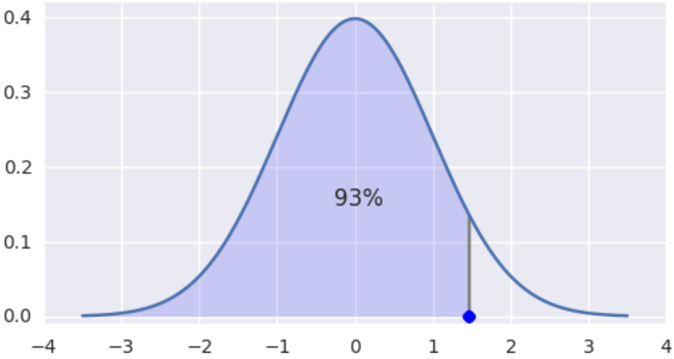
(C) For a normal distribution, IQR is less than $2 \times SD$.

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

Inverse Normal Distribution

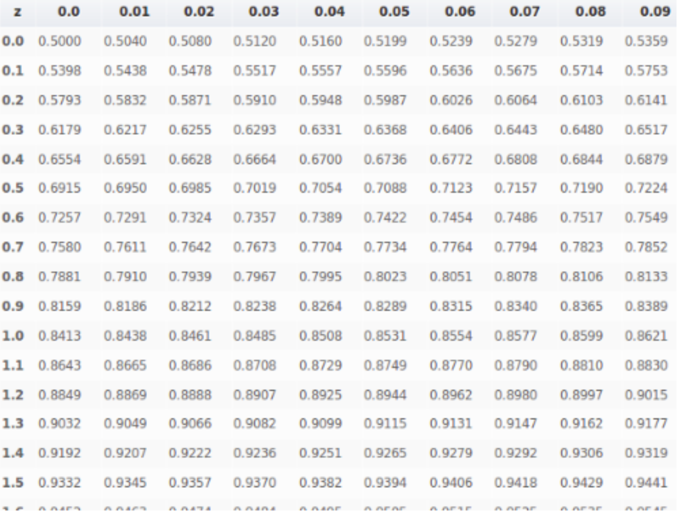
Finding the x Value for a Known Probability

Assume you are now given a probability value, *e.g.*, 93%, and you want to now which value of z^* is such that $p(z \leq z^*) = 0.93$



Inverse Normal Distribution

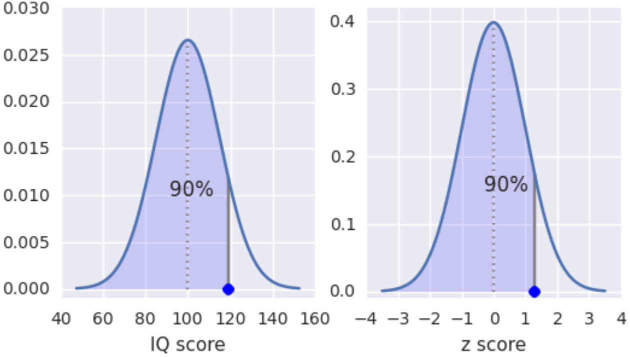
- We still use the tables of the normal distribution, but we now read the probability and find the row and column associated to 0.93.



z	0.0	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
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545

Practice

Given that the IQ score is normally distributed with $\mu = 100$ and $\sigma = 15$, find the IQ value for which only 10% of people have a higher score.



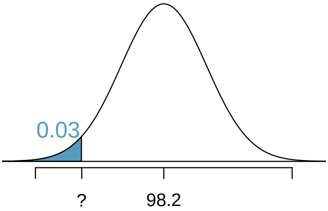
Remember: $z = \frac{x-\mu}{\sigma} \Rightarrow x = \mu + z\sigma \rightarrow$ `NORMSINV(0.9)`

Inverse Normal Distribution: Finding cutoff points

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

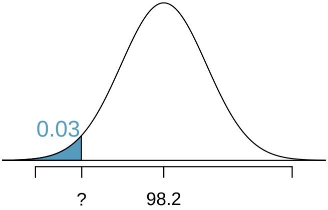
Inverse Normal Distribution: Finding cutoff points

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



Inverse Normal Distribution: Finding cutoff points

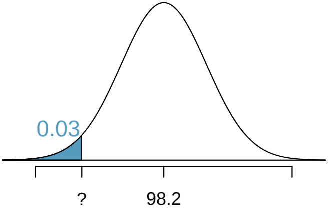
Body temperatures of healthy humans are distributed nearly normally with mean 98.2F and standard deviation 0.73F. What is the cutoff for the lowest 3% of human body temperatures?



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

Inverse Normal Distribution: Finding cutoff points

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

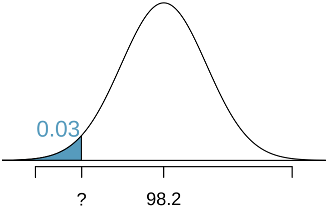


0.09	0.08	0.07	0.06	0.05	Z
0.0233	0.0239	0.0244	0.0250	0.0256	-1.9
0.0294	0.0301	0.0307	0.0314	0.0322	-1.8
0.0367	0.0375	0.0384	0.0392	0.0401	-1.7

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

Inverse Normal Distribution: Finding cutoff points

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

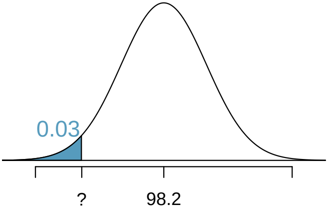


0.09	0.08	0.07	0.06	0.05	Z
0.0233	0.0239	0.0244	0.0250	0.0256	-1.9
0.0294	0.0301	0.0307	0.0314	0.0322	-1.8
0.0367	0.0375	0.0384	0.0392	0.0401	-1.7

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

Inverse Normal Distribution: Finding cutoff points

Body temperatures of healthy humans are distributed nearly normally with mean 98.2F and standard deviation 0.73F. 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

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

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 98.2F and standard deviation 0.73F. What is the cutoff for the highest 10% of human body temperatures?

(A) 97.3F

(B) 99.1F

(C) 99.4F

(D) 99.6F

Practice

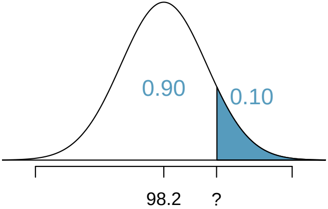
Body temperatures of healthy humans are distributed nearly normally with mean 98.2F and standard deviation 0.73F. What is the cutoff for the highest 10% of human body temperatures?

(A) 97.3F

(B) *99.1F*

(C) 99.4F

(D) 99.6F



Practice

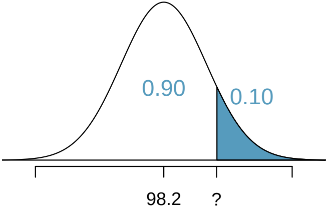
Body temperatures of healthy humans are distributed nearly normally with mean 98.2F and standard deviation 0.73F. What is the cutoff for the highest 10% of human body temperatures?

(A) 97.3F

(B) *99.1F*

(C) 99.4F

(D) 99.6F



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

Practice

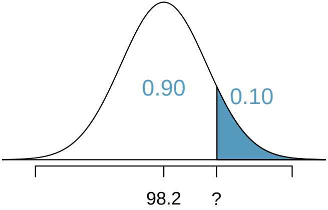
Body temperatures of healthy humans are distributed nearly normally with mean 98.2F and standard deviation 0.73F. What is the cutoff for the highest 10% of human body temperatures?

(A) 97.3F

(B) *99.1F*

(C) 99.4F

(D) 99.6F



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

$$P(X > x) \quad = \quad 0.10 \rightarrow P(Z < 1.28) = 0.90$$

Practice

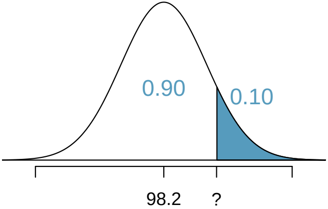
Body temperatures of healthy humans are distributed nearly normally with mean 98.2F and standard deviation 0.73F. What is the cutoff for the highest 10% of human body temperatures?

(A) 97.3F

(B) *99.1F*

(C) 99.4F

(D) 99.6F



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

$$P(X > x) = 0.10 \rightarrow P(Z < 1.28) = 0.90$$

$$Z = \frac{obs - mean}{SD} \rightarrow \frac{x - 98.2}{0.73} = 1.28$$

Practice

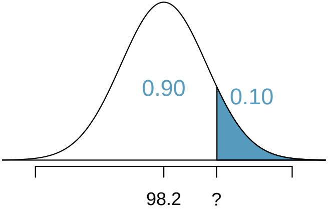
Body temperatures of healthy humans are distributed nearly normally with mean 98.2F and standard deviation 0.73F. What is the cutoff for the highest 10% of human body temperatures?

(a) 97.3F

(b) *99.1F*

(c) 99.4F

(d) 99.6F



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

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

1. Continuous Distributions

2. Uniform Distribution

3. Normal distribution
   - Normal distribution model
   - Standardizing with Z scores
   - Normal probability table
   - Normal probability examples
   - 68-95-99.7 rule

4. Evaluating the normal approximation
   - Normal probability plot

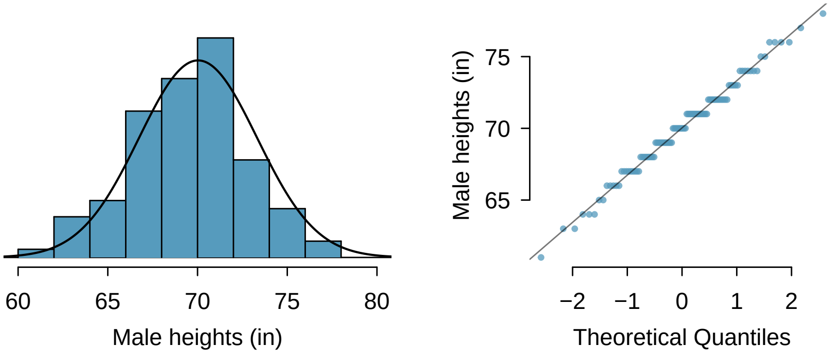
5. Normal Approximation to the Binomial

6. Exponential Distribution

7. Joint Cumulative Distribution Functions

Normal probability plot

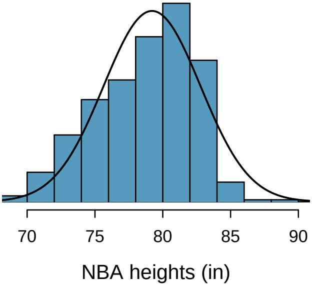
A histogram and normal probability plot of a sample of 100 male heights.

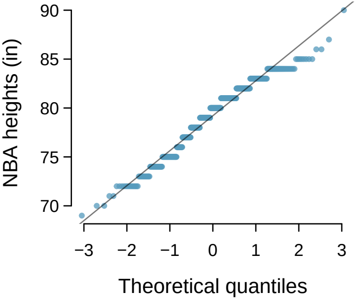


Anatomy of a normal probability plot

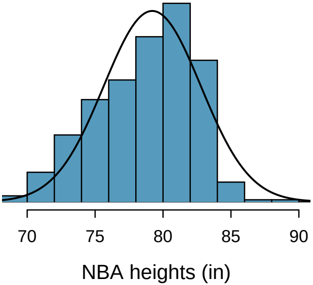
- Data are plotted on the y-axis of a normal probability plot, and theoretical quantiles (following a normal distribution) on the x-axis.
- If there is a linear relationship in the plot, then the data follow a nearly normal distribution.
- Constructing a normal probability plot requires calculating percentiles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

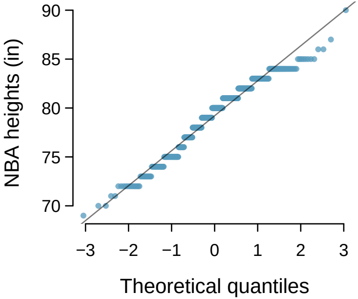
Below is a histogram and normal probability plot for the NBA heights from the 2008-2009 season. Do these data appear to follow a normal distribution?





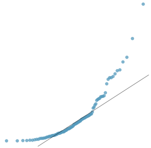
Below is a histogram and normal probability plot for the NBA heights from the 2008-2009 season. Do these data appear to follow a normal distribution?



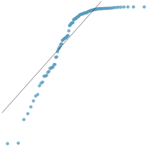


Why do the points on the normal probability have jumps?

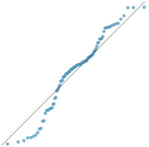
Normal probability plot and skewness



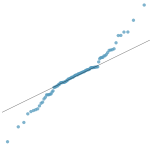
Right skew - Points bend up and to the left of the line.



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.

1. Continuous Distributions
2. Uniform Distribution
3. Normal distribution
   - Normal distribution model
   - Standardizing with Z scores
   - Normal probability table
   - Normal probability examples
   - 68-95-99.7 rule
4. Evaluating the normal approximation
   - Normal probability plot
5. **Normal Approximation to the Binomial**
6. Exponential Distribution
7. Joint Cumulative Distribution Functions

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

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.

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.

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.

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

This 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.

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

This seems like an awful lot of work...

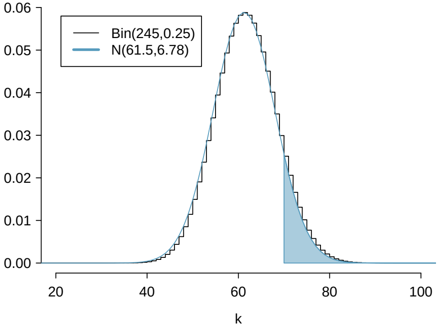
Normal approximation to the binomial

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 $p = 0.25$.

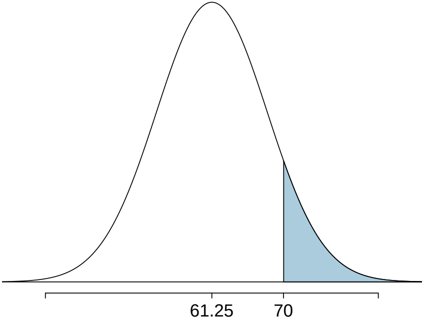
$$\mu = 245 \times 0.25 = 61.25 \qquad \sigma = \sqrt{245 \times 0.25 \times 0.75} = 6.78$$

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

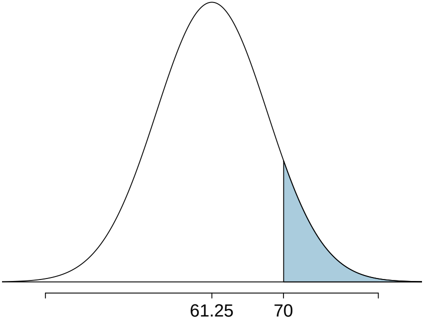


What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users?

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

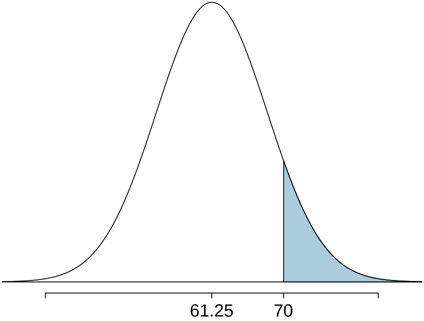


What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users?



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

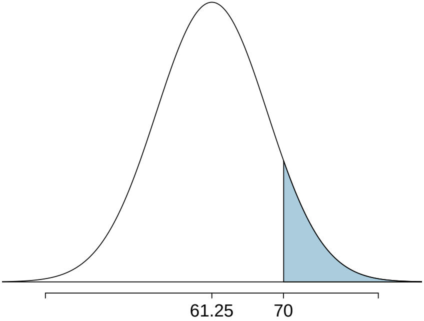
What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users?



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

	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

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

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

1. Continuous Distributions
2. Uniform Distribution
3. Normal distribution
   - Normal distribution model
   - Standardizing with Z scores
   - Normal probability table
   - Normal probability examples
   - 68-95-99.7 rule
4. Evaluating the normal approximation
   - Normal probability plot
5. Normal Approximation to the Binomial
6. **Exponential Distribution**
7. Joint Cumulative Distribution Functions

The Exponential Distribution

- Used to model the length of time between two occurrences of an event (the time between arrivals)
- Examples:
  - Time between trucks arriving at an unloading dock
  - Time between transactions at an ATM Machine
  - Time between phone calls to the main operator
- The exponential random variable $T(t > 0)$ has a probability density function

$$f(t) = \lambda e^{-\lambda t}, \quad t > 0$$

  - λ : number of occurrences in unit time
  - t : number of time units until the next occurrence
- The cumulative distribution function (the probability that an arrival time is less than some specified time t) is:

$$F(t) = 1 - e^{-\lambda t}$$

Practice

Customers arrive at the service counter at the rate of 15 per hour. What is the probability that the arrival time between consecutive customers is less than three minutes?

Practice

Customers arrive at the service counter at the rate of 15 per hour. What is the probability that the arrival time between consecutive customers is less than three minutes?

- The mean number of arrivals per hour is 15, so $\lambda = 15$

Practice

Customers arrive at the service counter at the rate of 15 per hour. What is the probability that the arrival time between consecutive customers is less than three minutes?

- The mean number of arrivals per hour is 15, so $\lambda = 15$
- Three minutes is .05 hours
- $p(\text{arrival time } < .05) = 1 - e^{-\lambda t} = 1 - e^{-(15)(.05)} = 0.5276$

Practice

Customers arrive at the service counter at the rate of 15 per hour. What is the probability that the arrival time between consecutive customers is less than three minutes?

- The mean number of arrivals per hour is 15, so $\lambda = 15$
- Three minutes is .05 hours
- $p(\text{arrival time } < .05) = 1 - e^{-\lambda t} = 1 - e^{-(15)(.05)} = 0.5276$
- Thus, there is a 52.76% probability that the arrival time between successive customers is less than three minutes

1 Continuous Distributions

2 Uniform Distribution

3 Normal distribution

Normal distribution model
Standardizing with Z scores
Normal probability table
Normal probability examples
68-95-99.7 rule

4 Evaluating the normal approximation

Normal probability plot

5 Normal Approximation to the Binomial

6 Exponential Distribution

7 Joint Cumulative Distribution Functions

Joint Cumulative Distribution Functions

- Let $X_1, X_2, \ldots, X_k$ be continuous random variables
- Their joint cumulative distribution function $F(X_1, X_2, \ldots, X_k)$ defines the probability that *simultaneously* X_1 is less than x_1, X_2 is less than x_2, etc., *i.e.*:

$$F(X_1, X_2, \ldots, X_k) = p(X_1 < x_1 \cap X_2 < x_2 \cap \ldots X_k < x_k)$$

- The random variables are independent if and only if:

$$F(X_1, X_2, \ldots, X_k) = F(X_1)F(X_2)\ldots F(X_k)$$

Covariance

- Let X and Y be continuous random variables, with means μ_x and μ_y
- The expected value of $(X - \mu_x)(Y - \mu_y)$ is called the covariance between X and Y:

$$Cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$$

- An alternative but equivalent expression is:

$$Cov(X, Y) = E(XY) - \mu_X \mu_y$$

- If the random variables X and Y are independent, then the covariance between them is 0. However, the converse is not true.

Sum of Random Variables

- Let $X_1, X_2, \ldots, X_k$ be continuous random variables with means $\mu_1, \mu_2, \ldots, \mu_k$ and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2$
- Then, the MEAN of their sum is the sum of their means, *i.e.*:

$$E(X_1 + X_2 + \ldots + X_k) = \mu_1 + \mu_2 + \ldots + \mu_k$$

- To compute the covariance:
  - If the covariance between every pair of these random variables is 0, then the VARIANCE of their sum is the sum of their variances:

$$Var(X_1 + X_2 + \ldots + X_k) = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_k^2$$

  - If the covariances between pairs of random variables are not 0, the VARIANCE of their sum is:

$$Var(X_1 + X_2 + \ldots + X_k) = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_k^2 + 2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} Cov(X_i, X_j)$$

Portfolio Analysis

- Consider two stocks, X and Y
  - The price of Stock X is normally distributed with mean 12 and standard deviation 4
  - The price of Stock Y is normally distributed with mean 20 and standard deviation 16
- The stock prices have a positive correlation, $\rho_{XY} = 0.50$
- Suppose you own 10 shares of Stock X and 30 shares of Stock Y

What is the probability that your portfolio value is less than €500?

Portfolio Analysis

- The mean and variance of this stock portfolio are:

$$\begin{aligned}
\mu_w &= 10\mu_X + 30\mu_Y = 10 \times 12 + 30 \times 20 = 720 \\
\sigma_w^2 &= 10^2\sigma_X^2 + 30^2\sigma_Y^2 + 2 \times 10 \times 30 \times Corr(X,Y) \times \sigma_X \times \sigma_Y \\
&= 10^2 \times 4^2 + 30^2 \times 16^2 + 2 \times 10 \times 30 \times 0.5 \times 4 \times 16 = 251,200 \\
\sigma_w &= \sqrt{251,200} = 501.20
\end{aligned}$$

- The z score for 500 is:

$$z = \frac{500 - 720}{501.20} = -0.44$$

- From the table, we get:

$$p(Z < -0.44) = 0.33$$

- Conclusion: There is a 33% probability that your portfolio value is less than €500.