Lecture 7: Random variables]

Statistics 101

Mine Çetinkaya-Rundel

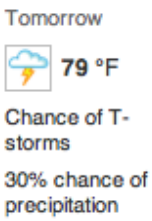
September 20, 2011

Review of commonly missed questions on the online quiz

OpenIntro quiz 2: questions 4 and 5

Let's bet...

According to wunderground.com there's *30%* chance of rain tonight in Durham. I'll bet you \$5 that it will rain tonight. That means, if it rains I win, and you give me \$5. If it doesn't rain, you win, and I give you \$5. What's *your* expected winnings from this game?



Event	Winnings	Probability	Winnings × Probability
Rain	−\$5	0.30	$-\$5 \times 0.30 = -1.5$
Doesn't rain	\$5	0.70	$\$5 \times 0.70 = \ \ 3.5$
			Total = \$2

↑ *probability model*

↑ *expected value* $E(X)$

Expected value and standard deviation

Let X = winnings.

X	$P(X)$	$X \times P(X)$	$(X - E(X))^2$	$(X - E(X))^2 \times P(X)$
−5	0.30	$-5 \times 0.30 = -1.5$	$(-5-2)^2 = 49$	$49 \times 0.30 = 14.7$
5	0.70	$5 \times 0.70 = 3.5$	$(5-2)^2 = 9$	$9 \times 0.70 = 6.3$
		$E(X) = 2$		$V(X) = 21$
				$SD(X) = \sqrt{21} = 4.58$

Your expected winnings is \$2, give or take \$4.58, i.e. if we played this game many times, on average you would win \$2, give or take \$4.58.

Should you play this game? Should I?

Expected value and standard deviation (cont.)

Expected value of a random variable

$$E(X) = \sum_i^n X_i \, P(X_i)$$

Standard deviation of a random variable

$$SD(X) = \sqrt{\sum_i^n (X_i - E(X))^2 \, P(X_i)}$$

Clicker question *(graded)*

A casino game costs $5 to play. If you draw first a red card, then the ace of hearts you win $500. If not, you don't win anything, i.e. lose your $5. What is your expected profits/losses from playing this game?
Remember: profit/loss = winnings - cost.

(a) A loss of 10¢
(b) A loss of 25¢
(c) A loss of 30¢
(d) A profit of 5¢

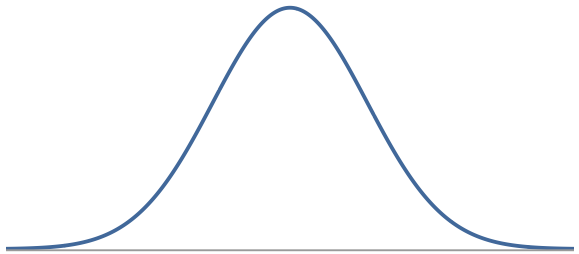
Fair game

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

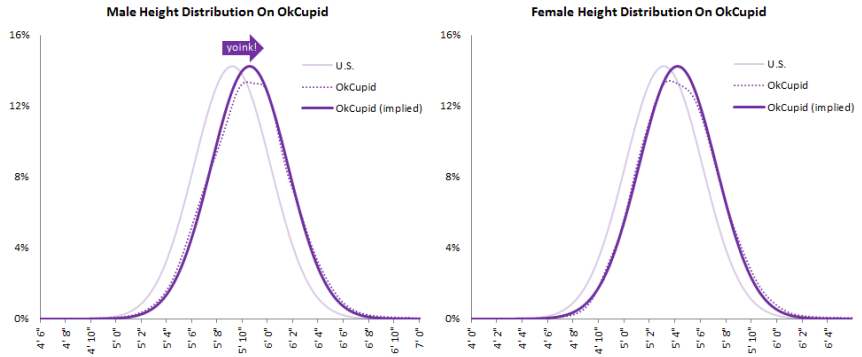
Do you think casino games in Vegas cost more or less than their expected payouts?

Normal distribution

- Unimodal and symmetric, bell shaped curve
- Most variables are nearly normal, but none are exactly normal
- Denoted as $N(\mu, \sigma)$



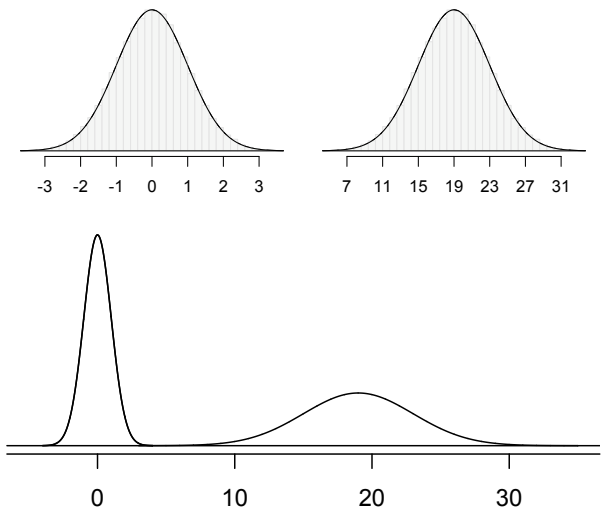
Heights of males and females



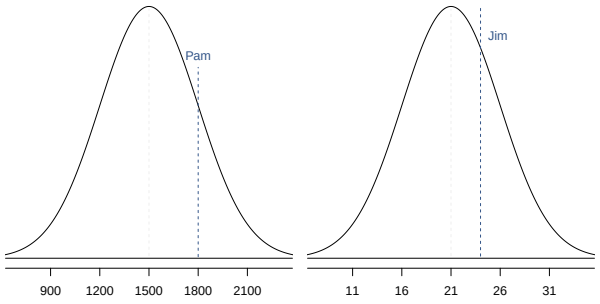
http://blog.okcupid.com/index.php/the-biggest-lies-in-online-dating/

Normal distributions with different parameters

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



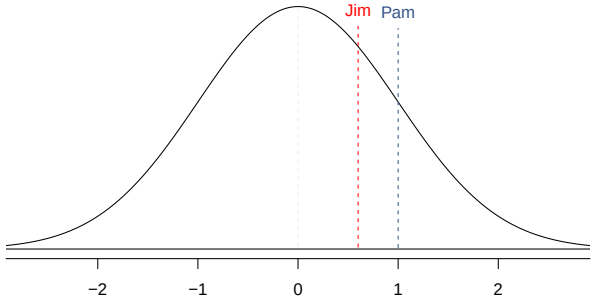
SAT scores are distributed nearly normally with mean 1500 and standard deviation 300. ACT scores are distributed nearly normally with mean 21 and standard deviation 5. A college admissions officer wants to determine which of the two applicants scored better on their standardized test with respect to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

- Since we cannot just compare these two raw scores, we should first convert them to *standardized* scores, i.e. *Z scores*.
- Z score of an observation is the number of standard deviations it falls above or below the mean.

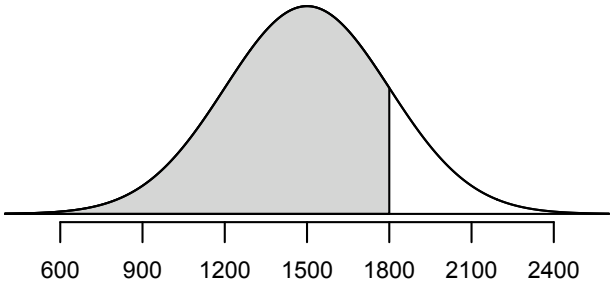
Z scores

$$Z = \frac{x - \mu}{\sigma}$$

Note: Z scores can be used to describe observations from distributions of any shape (not just normal) but only when the distribution is normal can we use Z scores to calculate percentiles.

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.



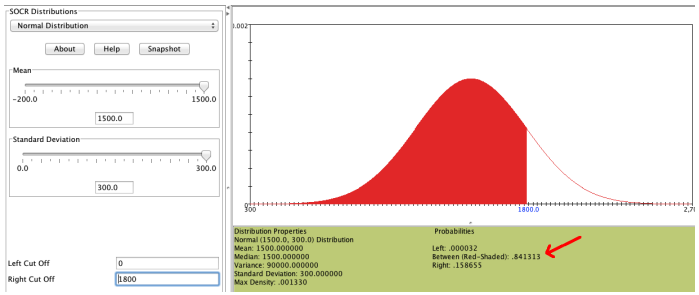
Calculating percentiles (computation)

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

- R:

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

- Applet: *http://www.socr.ucla.edu/htmls/SOCR_Distributions.html*

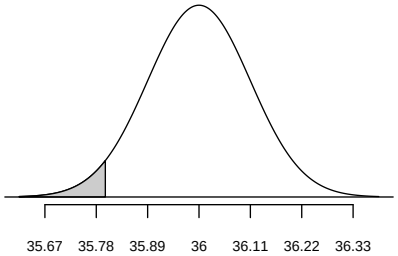


Calculating percentiles (tables)

	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

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 will be declared out of control. What's the probability that the amount of ketchup in a randomly selected bottle is less than 35.8 ounces?

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



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

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

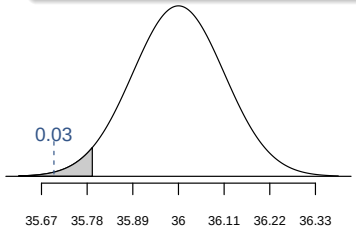
					Second decimal place of Z					
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	Z
0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179	−2.1
0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228	−2.0
0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287	−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

Clicker question

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 the bottle goes below 35.8 oz. or above 36.2 oz., then the bottle will be declared out of control. What percent of bottles are *not* declared out of control?

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

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. What is the cutoff for the lowest 3% of the amount of ketchup that goes in the bottles?



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

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

$$Z = \frac{x - \mu}{\sigma} \rightarrow \frac{x - 36}{0.11} = -1.88$$

$$x = (-1.88 * 0.11) + 36 = 35.7932$$

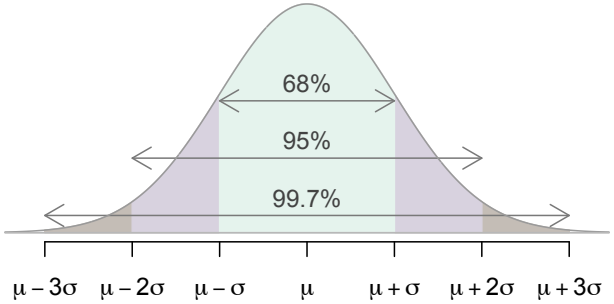
Clicker question

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. What is the cutoff for the highest 10% of the amount of ketchup that goes in the bottles?

(a) 35.86 oz
(b) 36.01 oz
(c) 36.09 oz
(d) 36.14 oz

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.

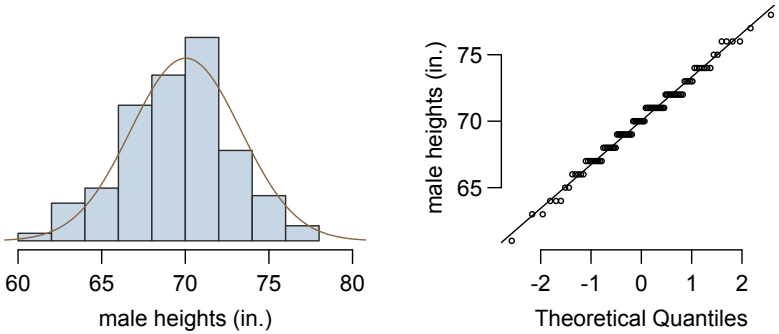


68-95-99.7 Rule (cont.)

- This rule is useful for
  - assessing normality, and
  - making quick estimates without using a Z table or a computer.
- For example to describe the amount of variability in the amount of ketchup in the bottles produced at the Heinz ketchup company, we could say:
  "95% of the bottles filled at the Heinz ketchup factory contain $36 - 2 * 0.11 = 35.78$ *oz* to $36 - 2 * 0.11 = 36.22$ *oz* ketchup."

Normal probability plot

- A histogram and *normal probability plot* of a sample of 100 male heights.
- The points appear to jump in increments in the normal probability plot since the observations are rounded to the nearest whole inch.



Anatomy of a normal probability plot

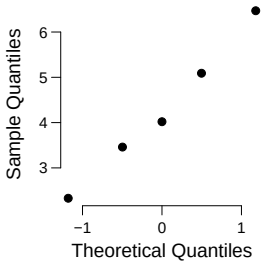
- Empirical quantiles (based on data) are plotted on the y-axis of a normal probability plot, and theoretical quantiles (following a normal distribution) on the x-axis.
- If there is a one-to-one relationship between the empirical and the theoretical quantiles, it means that the data follow a nearly normal distribution.
- Since a one-to-one relationship would appear as a straight line on a scatter plot, the closer the points are to a perfect straight line, the more confident we can be that the data follow the normal model.
- Constructing a normal probability plot requires calculating percentiles and corresponding z-scores for each observation, which is tedious. So we generally use software to construct these plots.

Construct a normal probability plot for the data set given below and determine if the data follow an approximately normal distribution.

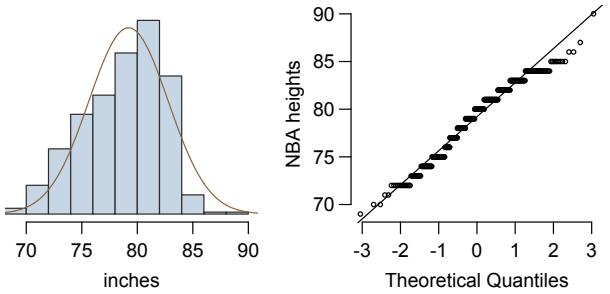
3.46, 4.02, 5.09, 2.33, 6.47

Observation i	1	2	3	4	5
x_i	2.33	3.46	4.02	5.09	6.47
Percentile = $\frac{i}{n+1}$	0.17	0.33	0.50	0.67	0.83
z_i	-0.95	-0.44	0	0.44	0.95

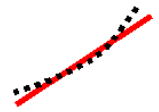
Since the points on the normal probability plot seem to follow a straight line we can say that the distribution is nearly normal.



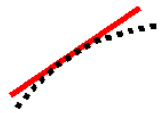
Below is a histogram and normal probability plot for the NBA heights from the 2008-9 season. Do these data appear to follow a normal distribution?



Normal probability plot and skewness

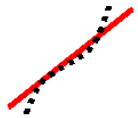


Right Skew - If the plotted points appear to bend up and to the left of the normal line that indicates a long tail to the right.



Left Skew - If the plotted points bend down and to the right of the normal line that indicates a long tail to the left.

Short Tails - An S shaped-curve indicates shorter than normal tails, i.e. narrower than expected.



Long Tails - A curve which starts below the normal line, bends to follow it, and ends above it indicates long tails. That is, you are seeing more variance than you would expect in a normal distribution, i.e. wider than expected.