HUMBEHV 3ST3

Normal Distribution & Hypothesis Testing

Prof. Patrick Bennett

Part 2 Sampling Distributions & the Normal Distribution

Sampling Distributions

- Most of our histograms show distributions of scores
- But sample statistics (e.g., mean) also have distributions
  - these are called sampling distributions
    - "higher-level" distributions of statistics, not scores
    - variation across samples, not individuals
  - every statistic (mean, standard deviation, r) has a sampling distribution
- Sampling distributions are IMPORTANT
  - used to make inferences about population parameters

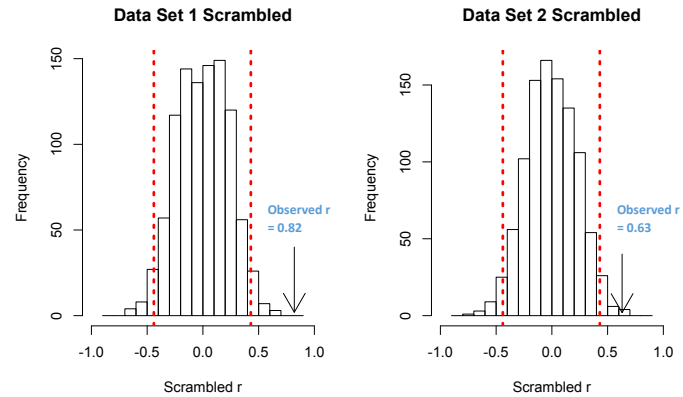
Example: Using Sampling Distributions of r

Histograms show distributions of r for scrambled (X,Y) data

Red dashed lines indicate boundaries which contain 95% of scrambled r's

r's outside boundaries are rare/unusual

Observed r's fall outside 95% boundaries, so they are unusual if we assume that the null hypothesis (pop r = 0) is true



Sampling Distributions & Inference

- sampling distributions are linked to statements about probability
- we need to know something about their shapes
- usually we do not know the shape of distributions of scores
  - so how can we know anything about shape of distribution of sample statistics like the mean?
- we use the **Central Limit Theorem**…

Sampling Distribution of the Mean

- Randomly select set of *n* scores X drawn from population
  - scores are distributed in population:
    - mean = *u*, variance = σ^2, shape of distribution = UNKNOWN
- Compute mean of sample & record for later use
- Repeat first 2 steps many times
- How is sample mean distributed? i.e., What is sampling distribution of mean?
  - In the long run, average mean will equal population mean (*u*)
  - In the long run, variance equals population variance (σ^2) divided by *n*
- Central Limit Theorem:
  - if sample *n* is sufficiently large, **the sampling distribution of the mean will be a normal distribution, regardless of the shape of the distribution of scores**

Central Limit Theorem

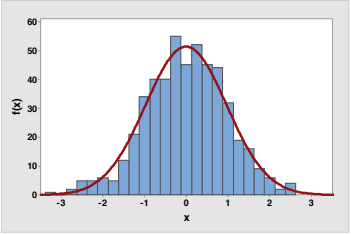
If a random variable *Y* has a population mean *u* and a population variance σ^2, then the sample mean, *M*, based on *n* observations, has an approximate normal distribution with mean *u* and a variance σ^2/n for sufficiently large *n*.

— B.S. Everitt, The Cambridge Dictionary of Statistics

The Normal Distribution

- Normal distribution
  - "Bell Curve"; Gaussian distribution
- used often in statistics
- unimodal & symmetrical around mean
  - defined by mean and standard deviation

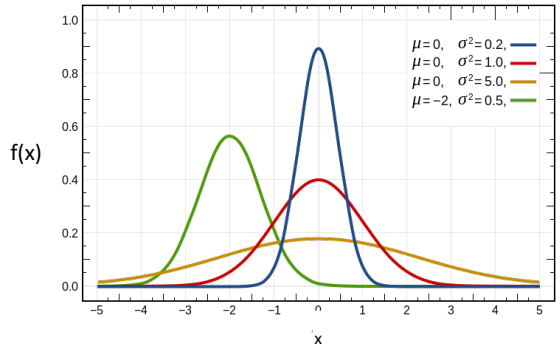
Carl Friedrich Gauss



histogram of 300 scores/observations selected randomly a normal distribution

red line is best-fitting Normal curve

Some Normal Distributions



Area under all normal curves — under all Probability Density Functions (PDFs) — always is 1.

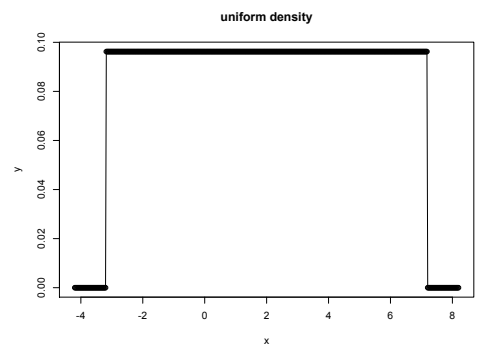
By Inductiveload - self-made, Mathematica, Inkscape, Public Domain, https://commons.wikimedia.org/w/index.php?curid=3817954

Why the Normal Distribution?

- Many variables we measure <u>are</u> normally distributed
- Many statistical procedures <u>assume</u> data are normally distributed
- Assuming a variable is normally distributed allows us to
  - describe the shape of the distribution
  - calculate probabilities for different values of variable
- **<u>Central Limit Theorem</u>** means that the assumption about normality often is reasonable

Central Limit Theorem Example

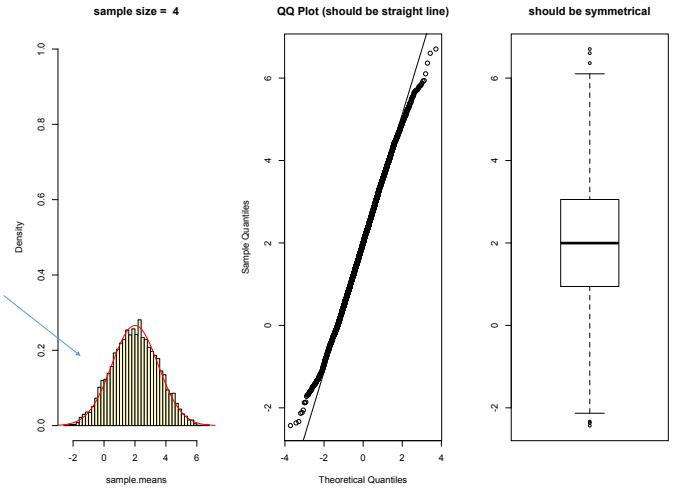
Uniform Distribution of <u>Scores</u>



Each sample consists of 4 <u>scores</u> selected randomly from uniform distribution ranging from -3.2 to +7.2

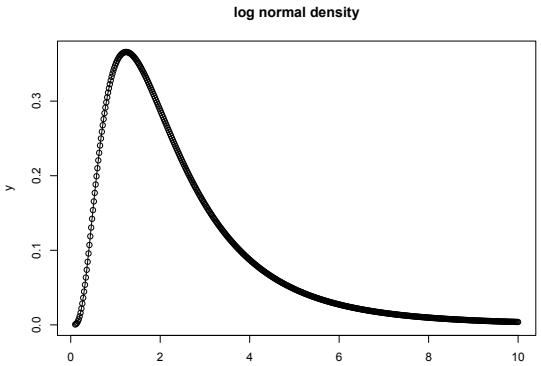
Means calculated for many samples (n=4). How are means distributed?

Sampling Distribution of the Mean is nearly Normal



41

Log-Normal Distribution of Scores

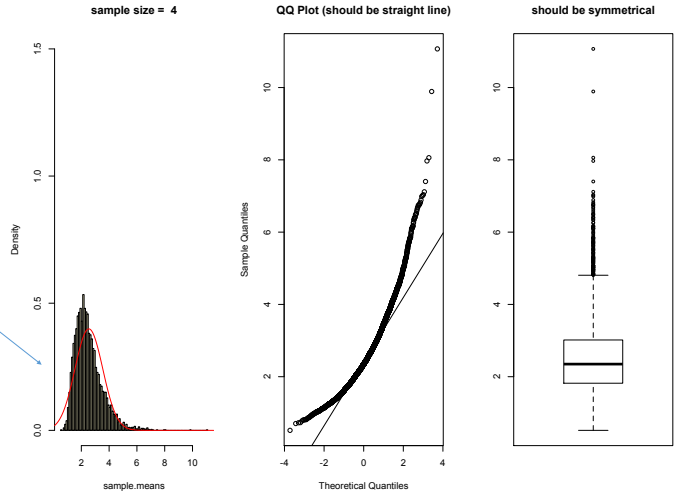


Each sample consists of N scores selected randomly from a log-normal distribution (positive skew)

Means calculated for many samples. How are means distributed?

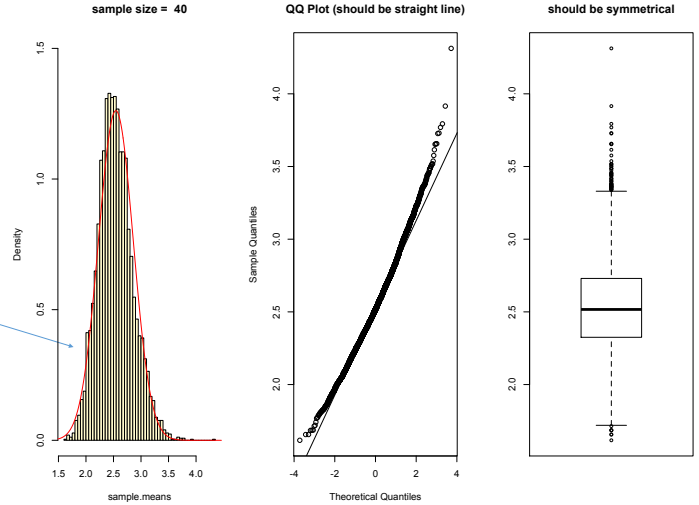
42

Sampling Distribution of the Mean is not Normal when N=4



43

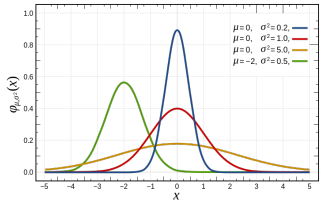
Sampling Distribution of the Mean is approximately Normal when N=40



44

z Transformation (Standard Scores)

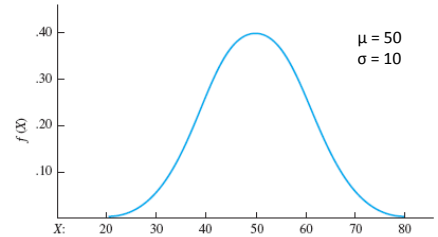
- Normal "distribution" is a family of distributions N(μ,σ²)
- We can convert <u>any</u> normal distribution into a <u>standard normal distribution</u> by transforming the values of our X into a standard normal score, z
- z scores are distributed as N(0,1)
  - Normal, mean = 0, variance=1



$$z = \frac{X - \mu}{\sigma}$$

Example: z transform

1. Subtract the value of the mean from each score.
   - 50 – 50 = 0
   - 40 – 50 = -10
   - And so on…
2. Divide each variable by σ.
   - 0/10 = 0
   - -10/10 = -1
   - +10/10 = +1
   - And so on…

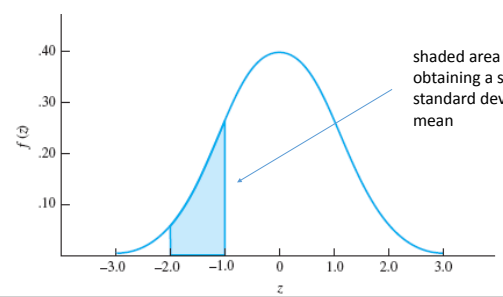


Z-scores

- If we transform our X_i values into Z_i values, these are called z-scores
- The value of z represents how many standard deviations, σ, an observation, i, is from the mean, μ
- Often use z-scores to compare distributions – similar to using *r* to compare linear trends between different pairs of (X,Y) variables

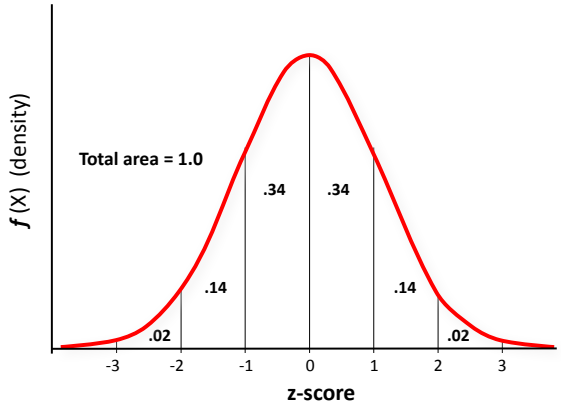
The standard normal distribution

- A normal distribution with mean (μ) = 0, and variance (σ²) = 1
- Denoted N(0, 1), where N stands for normal, zero is mean, and 1 is variance
- Use distribution to calculate probabilities:

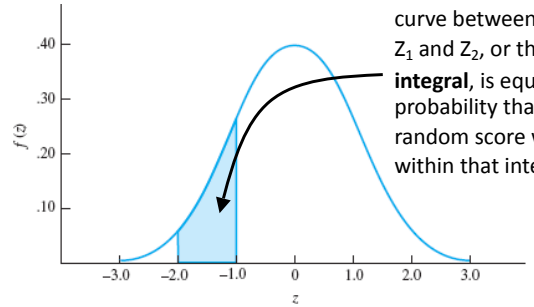


shaded area equals probability of obtaining a score between 1 & 2 standard deviations below the mean

Areas under the curve represent probabilities



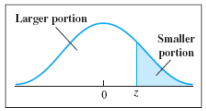
Finding area from a table



The area under the curve between points Z_1 and Z_2, or the **integral**, is equal to the probability that a random score will fall within that interval.

Instead of calculating the integral/area, we often use statistical tables.

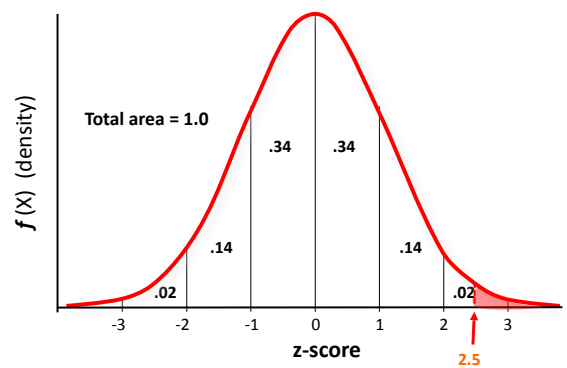
Table 6.1
The Normal Distribution (Abbreviated Version of Table E.10)



z	Mean to z	Larger Portion	Smaller Portion	z	Mean to z	Larger Portion	Smaller Portion
0.00	0.0000	0.5000	0.5000	0.45	0.1736	0.6736	0.3264
0.01	0.0040	0.5040	0.4960	0.46	0.1772	0.6772	0.3228
0.02	0.0080	0.5080	0.4920	0.47	0.1808	0.6808	0.3192
0.03	0.0120	0.5120	0.4880	0.48	0.1844	0.6844	0.3156
0.04	0.0160	0.5160	0.4840	0.49	0.1879	0.6879	0.3121
0.05	0.0199	0.5199	0.4801	0.50	0.1915	0.6915	0.3085
...	...	...	...	...	...	...	...
0.97	0.3340	0.8340	0.1660	1.42	0.4222	0.9222	0.0778
0.98	0.3365	0.8365	0.1635	1.43	0.4236	0.9236	0.0764
0.99	0.3389	0.8389	0.1611	1.44	0.4251	0.9251	0.0749
1.00	0.3413	0.8413	0.1587	1.45	0.4265	0.9265	0.0735
1.01	0.3438	0.8438	0.1562	1.46	0.4279	0.9279	0.0721
1.02	0.3461	0.8461	0.1539	1.47	0.4292	0.9292	0.0708
1.03	0.3485	0.8485	0.1515	1.48	0.4306	0.9306	0.0694
1.04	0.3508	0.8508	0.1492	1.49	0.4319	0.9319	0.0681
1.05	0.3531	0.8531	0.1469	1.50	0.4332	0.9332	0.0668
...	...	...	...	...	...	...	...

Making use of z-scores

- What is the probability of randomly selecting a score that is at least 2.5 standard deviations higher than the mean?



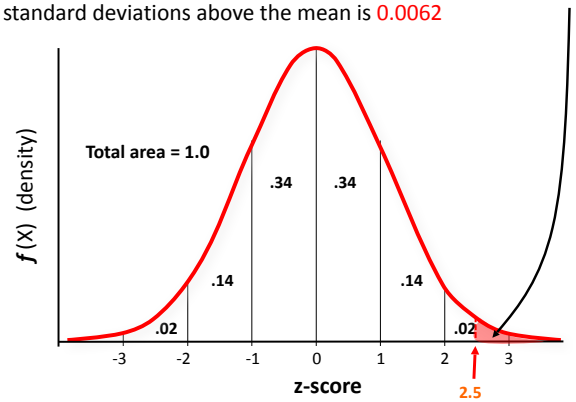
We consult our z-table…

Q. Which of these values gives us the answer we're looking for?

z	Mean to z	Larger Portion	Smaller Portion	z	Mean to z	Larger Portion	Smaller Portion
0.00	0.0000	0.5000	0.5000	0.45	0.1736	0.6736	0.3264
0.01	0.0040	0.5040	0.4960	0.46	0.1772	0.6772	0.3228
0.02	0.0080	0.5080	0.4920	0.47	0.1808	0.6808	0.3192
0.03	0.0120	0.5120	0.4880	0.48	0.1844	0.6844	0.3156
0.04	0.0160	0.5160	0.4840	0.49	0.1879	0.6879	0.3121
0.05	0.0199	0.5199	0.4801	0.50	0.1915	0.6915	0.3085
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
0.97	0.3340	0.8340	0.1660	1.42	0.4222	0.9222	0.0778
0.98	0.3365	0.8365	0.1635	1.43	0.4236	0.9236	0.0764
0.99	0.3389	0.8389	0.1611	1.44	0.4251	0.9251	0.0749
1.00	0.3413	0.8413	0.1587	1.45	0.4265	0.9265	0.0735
1.01	0.3438	0.8438	0.1562	1.46	0.4279	0.9279	0.0721
1.02	0.3461	0.8461	0.1539	1.47	0.4292	0.9292	0.0708
1.03	0.3485	0.8485	0.1515	1.48	0.4306	0.9306	0.0694
1.04	0.3508	0.8508	0.1492	1.49	0.4319	0.9319	0.0681
1.05	0.3531	0.8531	0.1469	1.50	0.4332	0.9332	0.0668
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
1.95	0.4744	0.9744	0.0256	2.40	0.4918	0.9918	0.0082
1.96	0.4750	0.9750	0.0250	2.41	0.4920	0.9920	0.0080
1.97	0.4756	0.9756	0.0244	2.42	0.4922	0.9922	0.0078
1.98	0.4761	0.9761	0.0239	2.43	0.4925	0.9925	0.0075
1.99	0.4767	0.9767	0.0233	2.44	0.4927	0.9927	0.0073
2.00	0.4772	0.9772	0.0228	2.45	0.4929	0.9929	0.0071
2.01	0.4778	0.9778	0.0222	2.46	0.4931	0.9931	0.0069
2.02	0.4783	0.9783	0.0217	2.47	0.4932	0.9932	0.0068
2.03	0.4788	0.9788	0.0212	2.48	0.4934	0.9934	0.0066
2.04	0.4793	0.9793	0.0207	2.49	0.4936	0.9936	0.0064
2.05	0.4798	0.9798	0.0202	2.50	0.4938	0.9938	0.0062

Making use of z-scores

- The probability of selecting a score that is more than 2.5 standard deviations above the mean is 0.0062

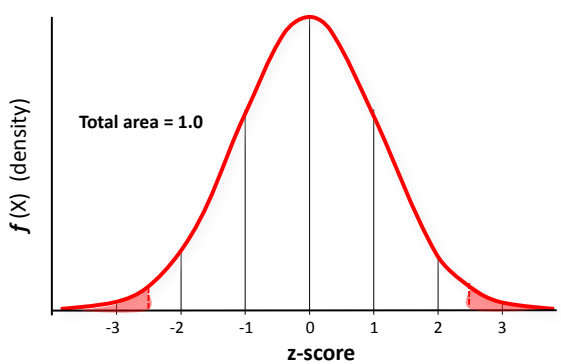


Q1. What is the probability of selecting a score more than 2.5 SDs ***below*** the mean?

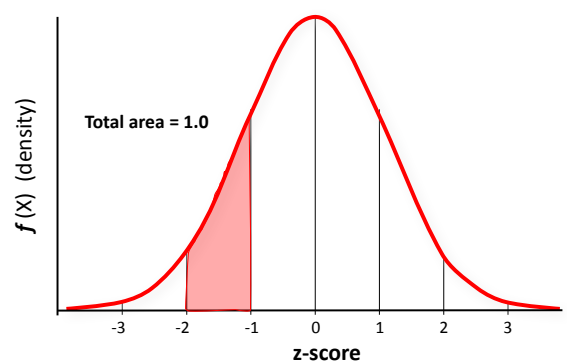
p(z < -2.5) = 0.006

Q2. What is the probability of getting a score more than 2.5 SDs ***above <u>or</u> below*** the mean?

p(z < -2.5) OR p(z > 2.5) = 0.006 + 0.006 = 0.012



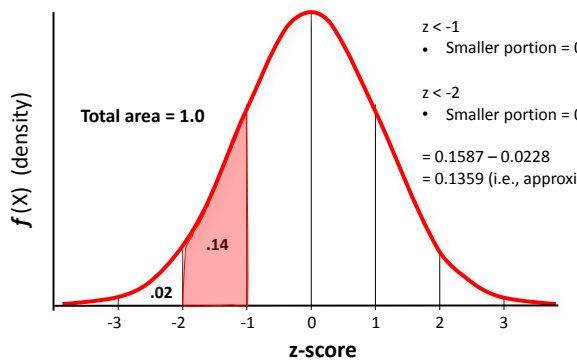
Q3. What is the probability of randomly selecting a score between -1 and -2 SDs from the mean?



We consult our z-table…

	z	Mean to z	Larger Portion	Smaller Portion	z	Mean to z	Larger Portion	Smaller Portion
	0.00	0.0000	0.5000	0.5000	0.45	0.1736	0.6736	0.3264
	0.01	0.0040	0.5040	0.4960	0.46	0.1772	0.6772	0.3228
	0.02	0.0080	0.5080	0.4920	0.47	0.1808	0.6808	0.3192
	0.03	0.0120	0.5120	0.4880	0.48	0.1844	0.6844	0.3156
	0.04	0.0160	0.5160	0.4840	0.49	0.1879	0.6879	0.3121
	0.05	0.0199	0.5199	0.4801	0.50	0.1915	0.6915	0.3085
	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
	0.97	0.3340	0.8340	0.1660	1.42	0.4222	0.9222	0.0778
	0.98	0.3365	0.8365	0.1635	1.43	0.4236	0.9236	0.0764
	0.99	0.3389	0.8389	0.1611	1.44	0.4251	0.9251	0.0749
p(z < -1) = 0.1587	1.00	0.3413	0.8413	0.1587	1.45	0.4265	0.9265	0.0735
	1.01	0.3438	0.8438	0.1562	1.46	0.4279	0.9279	0.0721
	1.02	0.3461	0.8461	0.1539	1.47	0.4292	0.9292	0.0708
	1.03	0.3485	0.8485	0.1515	1.48	0.4306	0.9306	0.0694
	1.04	0.3508	0.8508	0.1492	1.49	0.4319	0.9319	0.0681
	1.05	0.3531	0.8531	0.1469	1.50	0.4332	0.9332	0.0668
	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
	1.95	0.4744	0.9744	0.0256	2.40	0.4918	0.9918	0.0082
	1.96	0.4750	0.9750	0.0250	2.41	0.4920	0.9920	0.0080
	1.97	0.4756	0.9756	0.0244	2.42	0.4922	0.9922	0.0078
	1.98	0.4761	0.9761	0.0239	2.43	0.4925	0.9925	0.0075
	1.99	0.4767	0.9767	0.0233	2.44	0.4927	0.9927	0.0073
p(z < -2) = 0.0228	2.00	0.4772	0.9772	0.0228	2.45	0.4929	0.9929	0.0071
	2.01	0.4778	0.9778	0.0222	2.46	0.4931	0.9931	0.0069
	2.02	0.4783	0.9783	0.0217	2.47	0.4932	0.9932	0.0068
	2.03	0.4788	0.9788	0.0212	2.48	0.4934	0.9934	0.0066
	2.04	0.4793	0.9793	0.0207	2.49	0.4936	0.9936	0.0064
	2.05	0.4798	0.9798	0.0202	2.50	0.4938	0.9938	0.0062

Q3. What is the probability that a student would score between -1 and -2 standard deviations from the mean



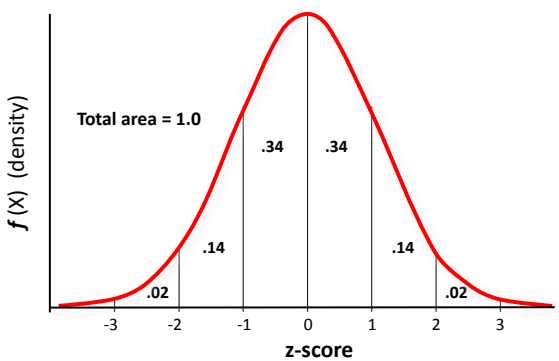
z < -1
- Smaller portion = 0.1587

z < -2
- Smaller portion = 0.0228

= 0.1587 – 0.0228
= 0.1359 (i.e., approximately 0.14)

Setting probable limits on an observation

What is the z-score cutoff for the lower 85% of the scores?



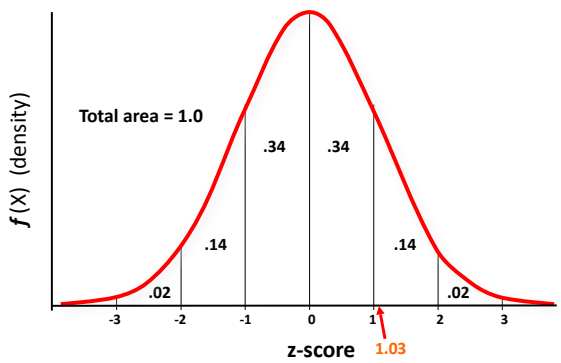
We consult our z-table…

z-score between 1.03 & 1.04 cuts off lower 85%

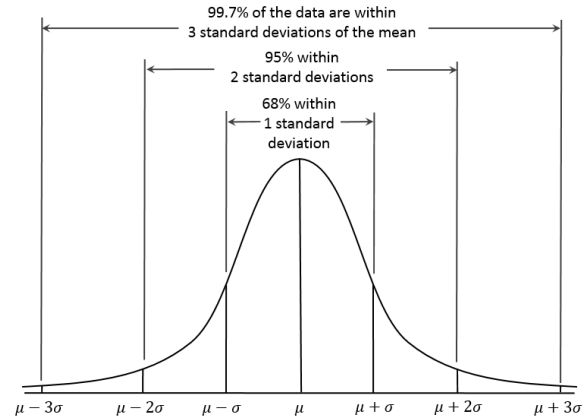
z	Mean to z	Larger Portion	Smaller Portion	z	Mean to z	Larger Portion	Smaller Portion
0.00	0.0000	0.5000	0.5000	0.45	0.1736	0.6736	0.3264
0.01	0.0040	0.5040	0.4960	0.46	0.1772	0.6772	0.3228
0.02	0.0080	0.5080	0.4920	0.47	0.1808	0.6808	0.3192
0.03	0.0120	0.5120	0.4880	0.48	0.1844	0.6844	0.3156
0.04	0.0160	0.5160	0.4840	0.49	0.1879	0.6879	0.3121
0.05	0.0199	0.5199	0.4801	0.50	0.1915	0.6915	0.3085
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
0.97	0.3340	0.8340	0.1660	1.42	0.4222	0.9222	0.0778
0.98	0.3365	0.8365	0.1635	1.43	0.4236	0.9236	0.0764
0.99	0.3389	0.8389	0.1611	1.44	0.4251	0.9251	0.0749
1.00	0.3413	0.8413	0.1587	1.45	0.4265	0.9265	0.0735
1.01	0.3438	0.8438	0.1562	1.46	0.4279	0.9279	0.0721
1.02	0.3461	0.8461	0.1539	1.47	0.4292	0.9292	0.0708
1.03	0.3485	0.8485	0.1515	1.48	0.4306	0.9306	0.0694
1.04	0.3508	0.8508	0.1492	1.49	0.4319	0.9319	0.0681
1.05	0.3531	0.8531	0.1469	1.50	0.4332	0.9332	0.0668
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
1.95	0.4744	0.9744	0.0256	2.40	0.4918	0.9918	0.0082
1.96	0.4750	0.9750	0.0250	2.41	0.4920	0.9920	0.0080
1.97	0.4756	0.9756	0.0244	2.42	0.4922	0.9922	0.0078
1.98	0.4761	0.9761	0.0239	2.43	0.4925	0.9925	0.0075
1.99	0.4767	0.9767	0.0233	2.44	0.4927	0.9927	0.0073
2.00	0.4772	0.9772	0.0228	2.45	0.4929	0.9929	0.0071
2.01	0.4778	0.9778	0.0222	2.46	0.4931	0.9931	0.0069
2.02	0.4783	0.9783	0.0217	2.47	0.4932	0.9932	0.0068
2.03	0.4788	0.9788	0.0212	2.48	0.4934	0.9934	0.0066
2.04	0.4793	0.9793	0.0207	2.49	0.4936	0.9936	0.0064
2.05	0.4798	0.9798	0.0202	2.50	0.4938	0.9938	0.0062

Setting probable limits on an observation

What is the z-score cutoff for the lower 85% of the scores?



Normal Distribution Rules of Thumb



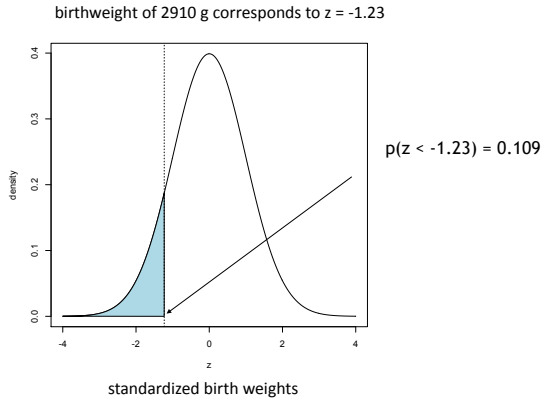
Using standardized scores (z test)

A woman in the US has just given birth to a full-term baby weighing 291 kg. Is this weight <u>unusually</u> low?

z test

- In US, <u>population</u> of Caucasian (non-Hispanic) full-term infants has a mean weight of 3480 g and a standard deviation of 462 g.
- The weights are <u>distributed</u> approximately normally.
- A weight of 2910 g is 1.23 standard deviations below the mean:
  - z = (2910 - 3480) / 462 = -1.23
- What is the probability of observing a weight that is at least this low?

z test



z test

- In US, population of Caucasian (non-Hispanic) full-term infants has a mean weight of 3480 g and a standard deviation of 462 g.
- The weights are distributed approximately normally.
- A weight of 2910 g is 1.23 standard deviations below the mean:
  - z = (2910 - 3480) / 462 = -1.23
- What is the probability of observing a weight that is at least this low?
  - p(z < -1.23) = 0.109