Recall: Data Description Basics

Any descriptive analysis should include the following information:

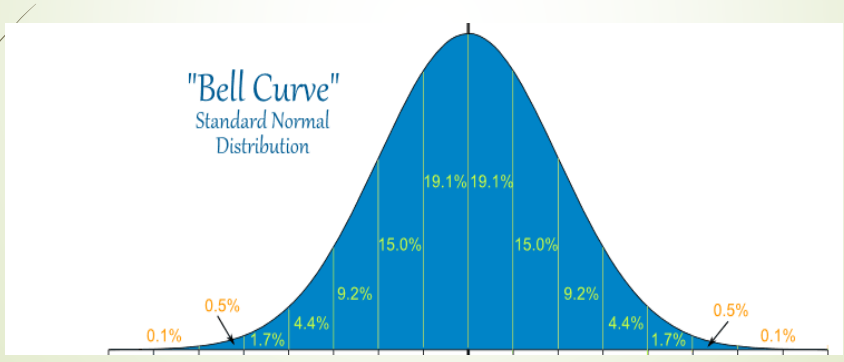
- A table that includes:
  - Appropriate metric for central tendency
  - Appropriate metrics for variability
  - Other relevant distribution metrics (skew or kurtosis)
  - Sample size, number of outliers
  - Results of a test for normality
- A figure of the data distribution (e.g. distribution curve, histogram or box plot)

1

Normality, Probability and Significance

Why did we spend all that time on normality?

- We use the standard **normal distribution curve** to determine the **probability** of a given value occurring for a standard normal population

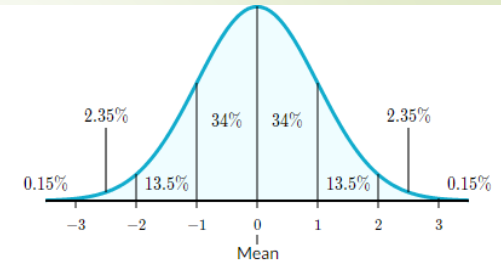


2

Focus on: The Normal Distribution Curve

The standard normal distribution is a continuous probability distribution that reflects the expected proportion of observations that should occur across all possible values of the variable of interest.

The area under the normal distribution curve represents the probability and the total area under the curve sums to one (all possible values that observations could take).



the three-sigma rule (or the 68-95-99.7 rule)

≈68, percent of the data falls within 1 standard deviation of the mean

≈95, percent of the data falls within 2 standard deviations of the mean

≈99.7 percent of the data falls within 3 standard deviations of the mean

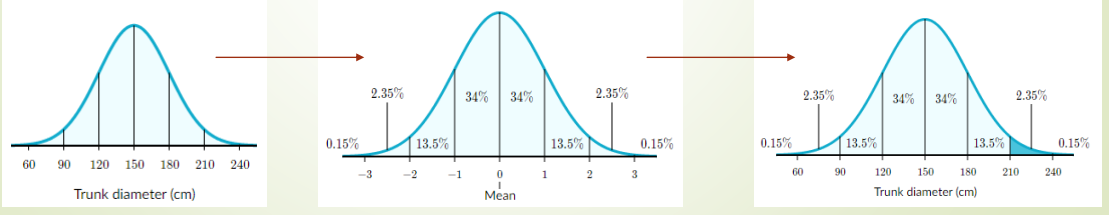
3

The Normal Distribution Curve

This standard normal curve differs from your actual data distribution curve in that its units are reported as a **z-score** that measure **how many standard deviations a given value is from the mean**.

We can use these z-scores to "translate" actual data values into a standard normal distribution that can be used to quantify probabilities.

In this way, **the standard normal distribution is how we get from measured values to probabilities**.



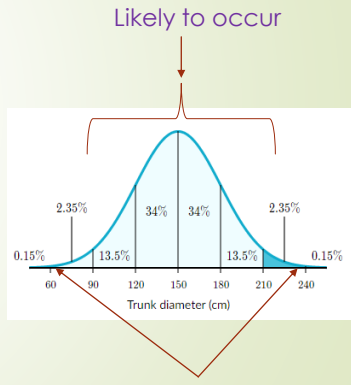
4

Focus on: Probabilities

Probabilities tell us how likely it is for a given value to occur. This forms the foundation for inferential statistical tests.

We use this probability to assess how likely it is that a result was achieved just due to random chance (fits the expected distribution).

- If the result is likely due to chance (probability > 0.05) we can conclude the result is not significant.
- If the result is not likely due to chance (probability < 0.5) it does not fit what is expected and we can conclude that the result is significant.



5

Focus on: Probabilities

- Statistical analyses are typically designed to test a specific research hypothesis

  For example:

  - There is a difference between two treatments
  - There is a linear relationship between two variables

But statistical tests actually quantify the probability that the null hypothesis is true

Null Hypothesis =

- The results you see are simply due to random chance.
- There is no significant difference, relationship, pattern, etc.

6

Focus on: Probabilities

- So, how do we go from data and a research hypothesis,
  - to a probability that the null hypothesis is true,
    - to a decision about the significance of a result and conclusion about our original hypothesis?

7

Hypothesis Testing

- All inferential tests start with a null hypothesis to test.
- Each type of null hypothesis has a specific formula that uses the data you have collected to quantify a **test statistic**.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left[\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\right]\left[\frac{n_1+n_2}{n_1 n_2}\right]}}$$

$$F = \frac{MST}{MSE}$$

$$MST = \frac{\sum_{i=1}^{k}(T_i^2/n_i) - G^2/n}{k-1}$$

$$MSE = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i=1}^{k}(T_i^2/n_i)}{n-k}$$

$$t = \frac{(\Sigma D)/N}{\sqrt{\frac{\Sigma D^2 - \frac{(\Sigma D)^2}{N}}{(N-1)(N)}}}$$

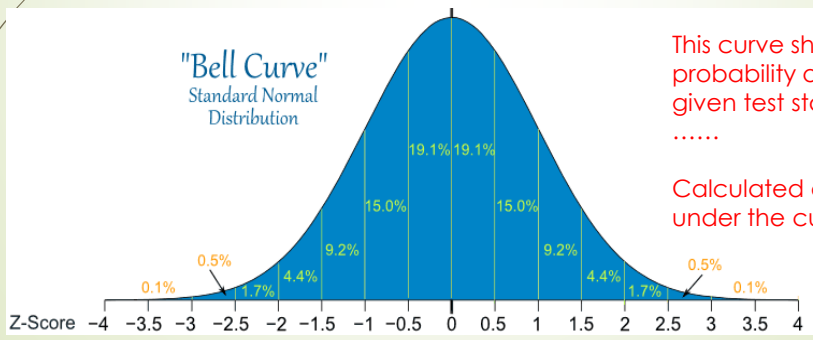
$$Z_o = \left| \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \right|$$

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]}}$$

8

Hypothesis Testing

- We then use the **normal distribution curve** to determine the **probability** of obtaining a test statistic that extreme IF THE NULL WERE TRUE.



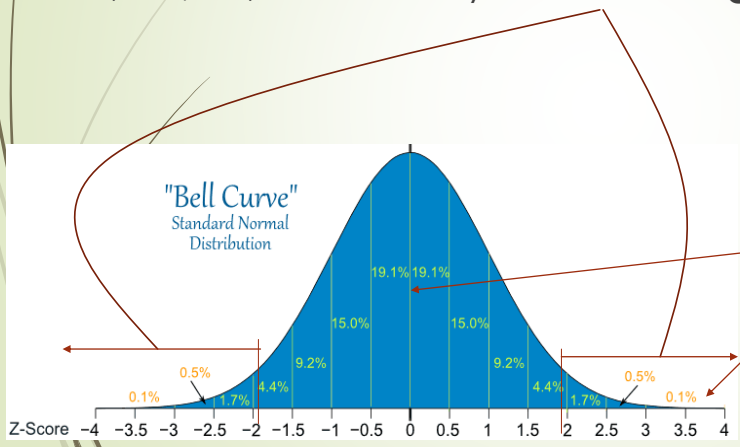
This curve shows the probability of obtaining a given test statistic value ……

Calculated as the area under the curve

9

Hypothesis Testing

- If the **probability** (area under the curve) of calculating a test statistic that extreme is less than our significance (alpha) threshold (usually 0.05), we can say we have a **significant result.**



The rest (95% in the middle) is likely due to random chance and is not significant

Remember how the normal distribution never touches 0 on the horizontal axis?

P-values can be tiny

10

Hypothesis Testing

- Another way of thinking about this:

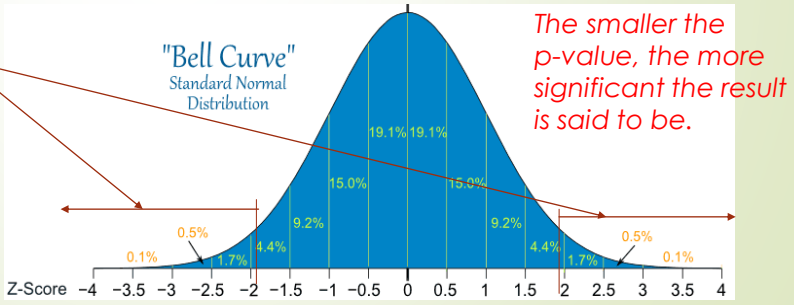
If the area under the curve (greater than or less than your calculated test statistic) is < 5 %

= Significant deviation from what is expected due to random chance

= reject the null

= we have a **significant result.**

= there is a difference, relationship, patterns, etc.



11

What is Significance?

- A statistical result is called **significant** if it is unlikely to have occurred by chance.
- Even though there is variability in the population, the probability of calculating a value this extreme simply due random variability is low (although not impossible)
- We use probabilities (**p-values**) and alpha threshold (usually $p<0.05$) to determine if a result is significant

12

What is Significance?

- Your alpha level is the degree of risk you are willing to take that you will reject a null hypothesis when it is actually true
- The probability that your result occurred by chance alone.
  - A p-value of 0.05 (5%) means that there is a 5% chance your result occurred by chance.
  - A p-value of 0.01 (1%) means that there is a 1% chance your result occurred by chance.
- Your chance of making a Type I error.

13

Focus – Types of Error

Probabilities are not absolute

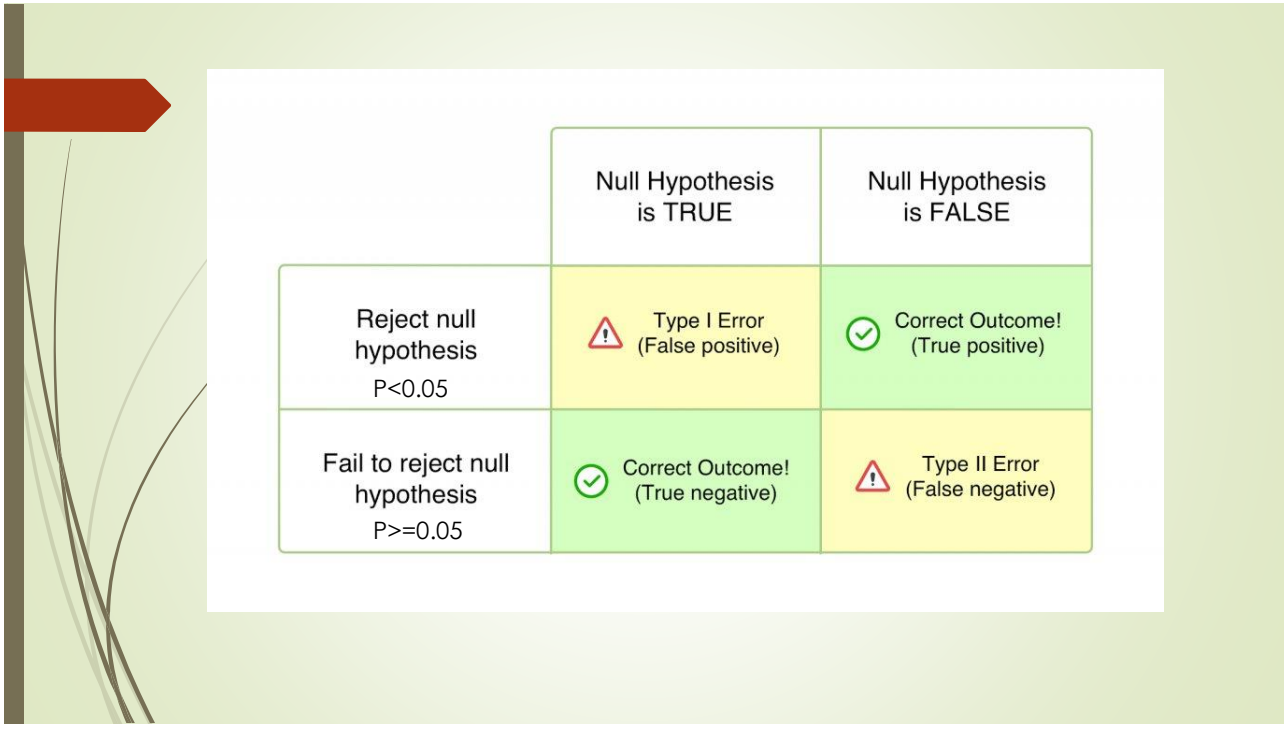
- Keep in mind probabilities are just that
- There is still a chance that we are incorrect
- **But HOW can we be wrong?**

We use the probability to determine significance

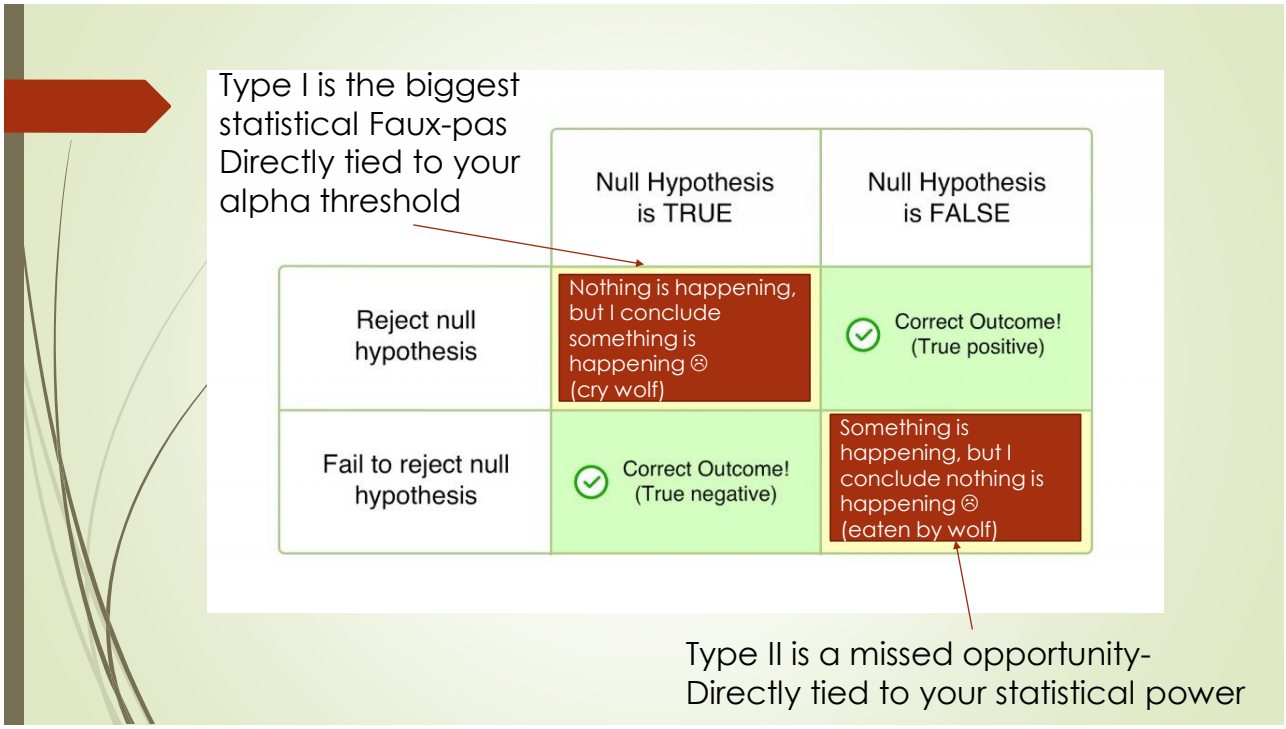
P-value (alpha threshold)

= the probability that you fit the normal distribution
= probability that the null is true
=risk of Type I error

14


Null Hypothesis is TRUE
Null Hypothesis is FALSE
Reject null hypothesis
P<0.05
Type I Error (False positive)
Correct Outcome! (True positive)
Fail to reject null hypothesis
P>=0.05
Correct Outcome! (True negative)
Type II Error (False negative)

15


Type I is the biggest statistical Faux-pas
Directly tied to your alpha threshold
Null Hypothesis is TRUE
Null Hypothesis is FALSE
Reject null hypothesis
Nothing is happening, but I conclude something is happening ☹ (cry wolf)
Correct Outcome! (True positive)
Fail to reject null hypothesis
Correct Outcome! (True negative)
Something is happening, but I conclude nothing is happening ☹ (eaten by wolf)
Type II is a missed opportunity-
Directly tied to your statistical power

16

Why p-value (alpha threshold) of 0.05?

"It is usual and convenient for experimenters to take 5 per cent as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard, and, by this means, to eliminate from further discussion the greater part of the fluctuations which chance causes have introduced into their experimental results."

Ronald Fisher

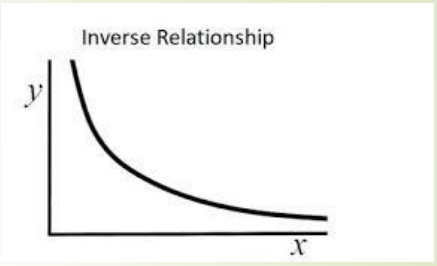
17

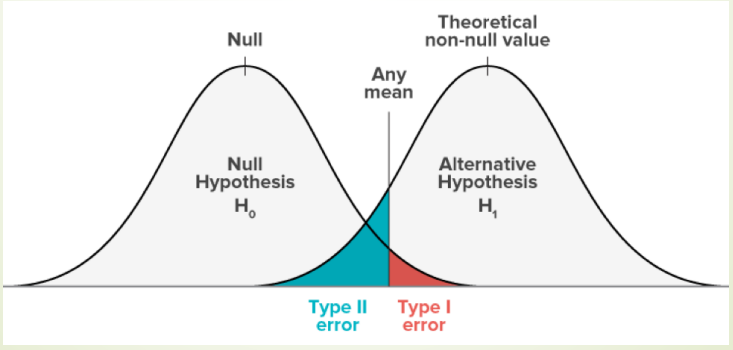
Don't we want to minimize the risk that we will reject a null hypothesis when it is actually true (minimize TYPE I error – false positive).

Why not an p-value (alpha threshold) of 0.01 (or smaller)?

18

There is an inverse relationship between the chance of Type I and Type II errors.
If you minimize your risk of a Type I error by changing the alpha threshold, you inherently increase the risk of a type II error.





19

Back to Probabilities

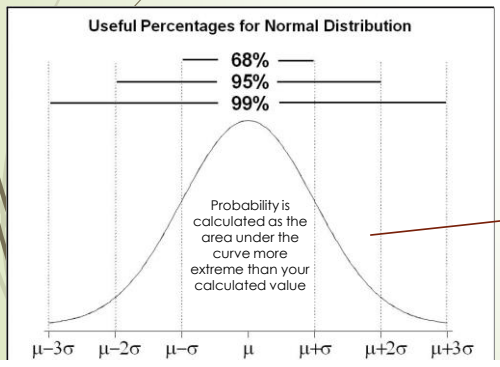
- So, how do we go from data and a research hypothesis,
  - to a probability that the null hypothesis is true,
    - to a decision about the significance of a result and conclusion about our original hypothesis?

20

To quantify a probability you first have to calculate a test statistic to locate on the normal probability curve.

- The normal curve can be considered your statistical **translator**.

It allows you to convert your data to a standardized test statistic with common units that can be assessed on a normal distribution curve



And then use that standardized test statistic value to quantify **probabilities** for a standard normal population

21

Z-scores: link **measured or hypothesized values** to **probabilities**

z Score (or standard score)

the number of standard deviations that a given value *x* is above or below the mean on a normal curve

$$z = \frac{(X - \bar{X})}{s},$$

22

Z-scores: link **measured or hypothesized values** to **probabilities**

When to use a z-score:

- Your sample data is normally distributed.
- You want to know how many standard deviations above or below the mean does a value lie.
- You want to use this to find the probabilities of different values occurring

Most common for the following applications:

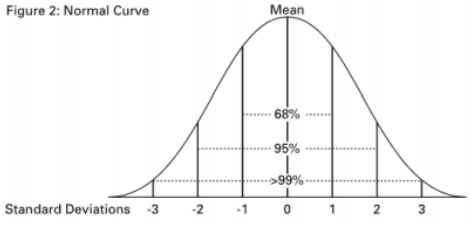
- Quantifying probabilities of interest
- Identifying outliers in a sample
- Determining in an observation belongs to a given population

23

Z-scores: linking observations to probabilities

Using z-scores for probability

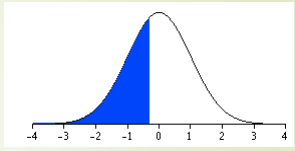
- So z-scores are essentially the x axis of the standard normal distribution
- They normalize any data set so that the mean = 0 and standard deviation = 1
- The areas under the curve tell you the probability of a certain z-score occurring.
- So we can use z-scores to determine probabilities.

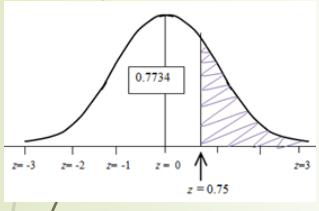


24

Finding Probabilities for z-scores

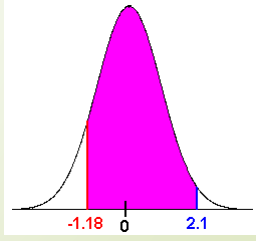
P(X< z)
denotes the probability of a value falling less than a given z-score (z)





P(*X* > z)
denotes the probability of a value falling above a given z score

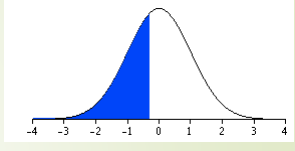
P(z^1 < *X* < z^2)
denotes the probability of a value falling between two different *z-scores*



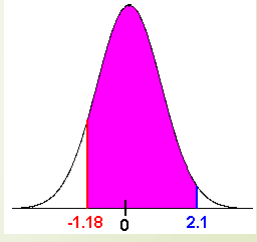
25

Finding Probabilities for z-scores

The following simple rules will help you find probabilities for any set of values:



- Examine the question carefully
- Sketch it out first!
- Calculate any z-score or X value you need
- Find associated probabilities
  - Look up Table: Know your flavor
  - Excel: always returns p<z
- Remember the rules:
  - The area under the curve always sums to 1
  - Both sides of the bell curve are symmetrical



26

Z-scores: linking observations to probabilities

Z-score and probability tips:

Use z to translate between measured values and probabilities $z = \frac{(X - \bar{X})}{s}$

If using a table:

Use a standard lookup table to find probabilities associated with a given z score

- Probabilities are in the body of the table
- Z-scores determine the rows and columns
- *Know the "flavor" of your table Most are P(X<z)*

Standard Normal Probabilities

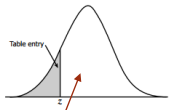


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

27

Z-scores: linking observations to probabilities

Z-score and probability tips:

Use z to translate between measured values and probabilities $z = \frac{(X - \bar{X})}{s}$

Or Use excel:

=normsdist(z) to return a probability

	A	B
1	=normsdist(0.76)	0.77637

=normsinv(p) to return a z-score for a given probability

	A	B
1	=normsinv(0.95)	1.64485

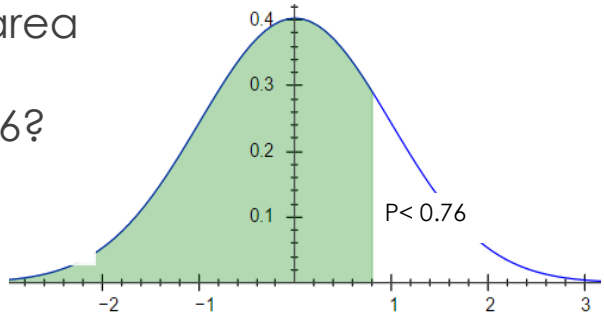
28

Examples of getting probabilities from z-scores

- Examine the question carefully
- Sketch it out first!
- Calculate any z-score or X value you need
- Find associated probabilities
  - Look up Table: Know your flavor
  - Excel: always returns p<z
- Remember the rules:
  - The area under the curve always sums to 1
  - Both sides of the bell curve are symmetrical

What percent of the area under the curve falls below a z-score of 0.76?

P (z< 0.76)

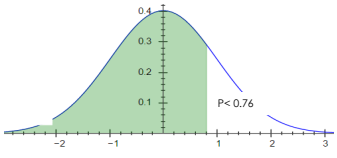


29

Examples of getting probabilities from z-scores

- Examine the question carefully
- Sketch it out first!
- Calculate any z-score or X value you need
- Find associated probabilities
  - Look up Table: Know your flavor
  - Excel: always returns p<z
- Remember the rules:
  - The area under the curve always sums to 1
  - Both sides of the bell curve are symmetrical

What percent of the area under the curve falls below a z-score of 0.76?



	A	B
1	=normsdist(0.76)	0.77637

~78% of observations fall below a z-score of 0.76

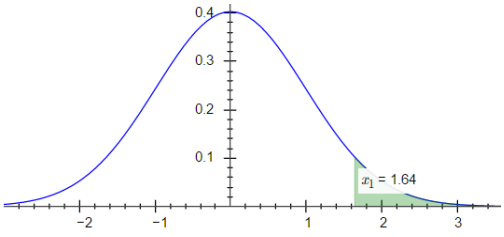
30

Examples of getting z-scores from probabilities

- Examine the question carefully
- Sketch it out first!
- Calculate any z-score or X value you need
- Find associated probabilities
  - Look up Table: Know your flavor
  - Excel: always returns p<z
- Remember the rules:
  - The area under the curve always sums to 1
  - Both sides of the bell curve are symmetrical

What is the z-score beyond which only 5% of all possible outcomes are higher?

z where P>z = 0.05



	A	B
1	=normsinv(0.95)	1.64485

Z where 5% of all observations are higher = 1.65

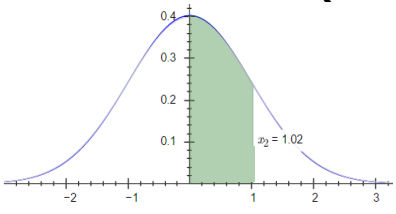
31

Examples of getting probabilities from data

- Examine the question carefully
- Sketch it out first!
- Calculate any z-score or X value you need
- Find associated probabilities
  - Look up Table: Know your flavor
  - Excel: always returns p<z
- Remember the rules:
  - The area under the curve always sums to 1
  - Both sides of the bell curve are symmetrical

In a distribution with a mean of 100 and a standard deviation of 15, what is the probability that a score will fall between 100 and 115 ?

P (100 >x>115)



Before you can find any probabilities you have to find z-scores

$$z = \frac{(X - \bar{X})}{s}$$

Z = (100 – 100) / 15 = 0
Z = (115 – 100) / 15 = 1

32

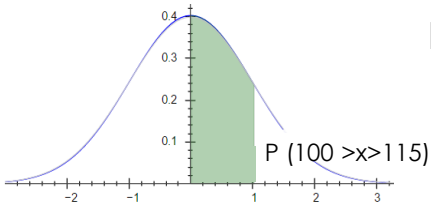
Examples of getting probabilities from data

- Examine the question carefully
- Sketch it out first!
- Calculate any z-score or X value you need
- Find associated probabilities
  - Look up Table: Know your flavor
  - Excel: always returns p<z
- Remember the rules:
  - The area under the curve always sums to 1
  - Both sides of the bell curve are symmetrical

In a distribution with a mean of 100 and a standard deviation of 15, what is the probability that a score will fall between 100 and 115 ?

P (100 >x>115)

P (0 >z>1)



	A	B
1	=normsdist(0)	0.50
2	=normsdist(1)	0.84

Take the difference:
0.84 – 0.5 = 0.34

Probability of observations between 100 and 115 ~ 34%

33

z-Scores in Practice

PLAYING WITH KITTIES

The life span of domesticated cats is normally distributed with a mean of 15.7 years and a standard deviation of 1.6 years.

Alison's cat is 18 years old. What is the probability that a cat will live to be as old as Alison's cat?

34

z-Scores in Practice

PLAYING WITH KITTIES

The life span of domesticated cats is normally distributed with a mean of 15.7 years and a standard deviation of 1.6 years.

Alison's cat is 18 years old. What is the probability that a cat will live to be as old as Alison's cat?

$$1 - P(z < 1.43)$$
$$= 1 - 0.9236$$
$$= 0.0764$$

= a 7.6% probability

35

Finding the value associated with a probability

- Sometimes you may want to do things the other way around:

  i.e. What values are associated with a less than 5% chance of occurring

  Useful for identifying outlier thresholds

36

z-Scores in Practice

PLAYING WITH KITTIES

The life span of domesticated cats is normally distributed with a mean of 15.7 years and a standard deviation of 1.6 years.

How old would Alison's cat have to be in order to be considered significantly old (use $p < 0.05$)

37

z-Scores in Practice

PLAYING WITH KITTIES

The life span of domesticated cats is normally distributed with a mean of 15.7 years and a standard deviation of 1.6 years.

How old would Alison's cat have to be in order to be considered significantly old (use $p < 0.05$)

1. *What do I want to know:*

   X where p x>z = 0.05

2. *What is the z-score associated with that probability?*

   =NORMSINV(0.95) = 1.644853627

38

z-Scores in Practice

PLAYING WITH KITTIES

The life span of domesticated cats is normally distributed with a mean of 15.7 years and a standard deviation of 1.6 years.

How old would Alison's cat have to be in order to be considered significantly old (use $p < 0.05$)

3. What is the x value associated with that z-score?

$$z = \frac{(X - \bar{X})}{s}$$

becomes

$X = (z * s) + \bar{x}$

x = (1.645*1.6)+15.7 = 18.432

18.4 years old is significantly old

39

Real Life Problems

The EPA wants to set drinking water standards. This is based on a cost benefit analysis that looks at the health impacts AND the economic impacts of implementation.

If they set the threshold for lead at **1** ppm what proportion of public buildings will require lead remediation?

Key metrics on the distribution of lead contamination in public building infrastructure:

Mean = 0.6

s = 0.2

40

What do I want to know? :

P(X>1)

What is the z-score associated with this value?:

$$z = \frac{(X - \bar{X})}{s}$$

z= (1-0.6)/0.2 = 2.00

What is probability associated with this z-score (p(X>1)?:

=NORMSDIST(2) = 0.98

1-0.98 ~0.02 or 2%

41

Real Life Problems

Chris needs a new ballroom dance partner.

At 6' 1" he insists that she must be between 5'8 and 5'11.

He is having trouble finding a woman who fits this criteria. He wants to know if he should give up.

What is the probability of Chris finding a woman between 5'8 and 5'11?

Mean = 65"

s = 2.7"

~0.12 or 12%

42