# MTH499/599 Lecture Notes 03

Donghui Yan

University of Massachusetts Dartmouth

# Outline

- Introduction to hypothesis testing
- t-test
- 2-sample test

# Reasoning under uncertainty

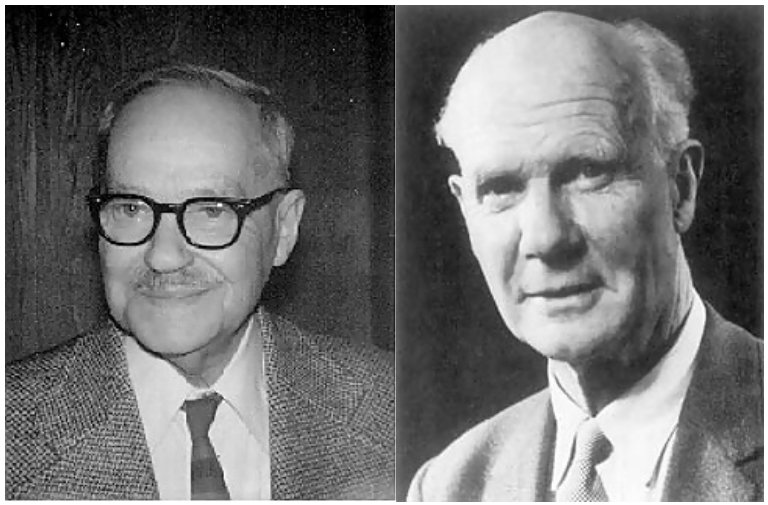
- We often need to draw conclusions
  - e.g., $A > B$ and $B > C$, so $A > C$
  - e.g., if one cannot pronounce *Lowell* correctly, there is good chance that he is new to MA
  - e.g., a good programmer would do well in data science
- Characteristics
  - Based on logical deduction
    - *Deterministic* and holds once and for all
  - Based on experience or other knowledge
    - *Uncertainty* involved
    - How to incorporate chance variation when drawing a conclusion?
    - ♠ Hypothesis testing provides such a framework.

# Why hypothesis testing in data science?

- Some may have learned hypothesis testing in elementary statistics
  - Often appears in statistics texts or scientific papers
- Hypothesis testing is an essential element of data science
  - We often draw conclusions by analyzing data
    - Substantial uncertainty involved
    - But we wish to draw conclusions in a sound framework
  - Many tasks in data science requires A/B test
    - Hypothesis testing allows to rigorously compare alternatives.

# Hypothesis testing

- Proposed by Neyman and Pearson in 1933
- Arguably the most important work in modern statistics
- A statistical framework for reasoning (drawing conclusions) under uncertainty
  - Follow a similar line of argument as *proof by contradiction*
  - Incorporate uncertainty
  - Quantify chance error.



# Proof by contradiction revisited

- Idea (steps)
  - *Assume what needs to proof is false*
  - *Carry out logical deduction and reach some contradiction*
  - *Conclude that the assumption is false*
- E.g., prove that $\sqrt{2}$ is an irrational number
  - Assume otherwise, i.e., $\sqrt{2}$ is rational
  - Can write $\sqrt{2} = a/b$ for $a, b \in \mathbb{N}$ and $(a, b) = 1$
    - $a^2 = 2b^2$, thus $2 \mid a$
    - Similarly, $b^2 = 2k^2$ and $2 \mid b$
    - Thus $2 \mid (a, b)$, a contradiction!
    - So assumption is invalid and $\sqrt{2}$ is irrational.

# Implementation of hypothesis

- Key idea
  - Find a way to evaluate the collected data sample
  - Expect to see "surprise" (rare or extreme events)
- Need to fix a quantity to look at
  - E.g., population mean $\mu$ (unknown)
  - Corresponding testing stat, $\overline{X}$
- A precise definition of "surprise"
  - How the testing statistics varies $\Rightarrow$ distribution
  - A cutoff value
    - How extreme a testing stat constitutes a "surprise".

# Steps in hypothesis testing

- Formulate null ($H_0$) and alternative ($H_a$) hypothesis
- Specify a significance level $\alpha$ (e.g., $\alpha = 0.05$)
  - Chance error in conclusion drawn would be at most $\alpha$
- Fix a *testing statistic* $T$ and determine *rejection region* $\mathcal{R}$
  - $\mathcal{R}$ often contains "extreme" values
- Collect a data sample $\boldsymbol{S}$ and calculate the value of $T$ on $\boldsymbol{S}$
  - If $T(\boldsymbol{S}) \in \mathcal{R}$, then *reject* null hypothesis $H_0$ at level $\alpha$
  - Otherwise *do not reject* $H_0$.

# Hypothesis testing Vs proof by contradiction

| | **Hypothesis testing** | **Proof by contradiction** |
|---|---|---|
| Assumption | $H_0$ is true | Conclusion to prove is false |
| Action | Collect data sample | Logical deduction |
| Expect | Small chance event (getting extreme values) | Logic statement contradictory to known facts or conditions |
| Decision | Rejection of $H_0$ | Rejection of assumption |
| Correctness | Possibility of chance error | Deterministic and always true |

- *Hypothesis testing* $\leftrightarrow$ *proof by contradiction*
  - Ooccurrence of small chance event $\leftrightarrow$ contradictory logical statement.

# Hypothesis testing as framework for drawing conclusion

- Mainly refer to Qs that can be formulated by a parameter of the population
  - e.g., population mean $\mu$, or proportion etc
    - Proportion can be reduced to mean
  - e.g., simple function of the population
    - Regression parameters (to be discussed later)
  - So $H_0$ and testing statistic all related to this parameter
- More general framework based on decision theory
  - Statistical decisions all attempt to minimize some "errors"
  - Hypothesis testing as a special case.

# Hypothesis testing

- Hypothesis testing (t-test) by example
- A few concepts and terminology

# An example on hypothesis testing

*Alex accidentally hit a scale with a hammer. Does the scale still work? As a Data Scientist, Alex figured out a way via hypothesis testing.*

*1). He used the broken scale to get 10 packs of sugar all at 50 grams (such weights measured by the broken scale).*

*2). Then he used a working scale and re-measured the weight of the 10 packs, which are*

*49, 50, 47.5, 49.6, 48.6, 49.0, 48.6, 49, 49.4, 50.2*

*He then carried out hypothesis testing at* $\alpha = 0.02$.

- Does the broken scale work (at $\alpha = 0.02$)?

# Example (2-side t-test)

a). What would be a suitable null hypothesis $H_0$ here?

b). Formulate a testing statistic T, and indicate its distribution.

c). For $\alpha = 0.02$, determine the rejection region.

d). Using the give data, calculate the value of T, and the corresponding p-value.

e). Draw conclusion under significance level $\alpha = 0.02$.

# Example (continued)

Here $\alpha = 0.02$. Null hypothesis $H_0 : \mu = 50$, $H_A : \mu \neq 50$.

- *Think of all sugar packs measured to be 50g by the broken scale as a population, the population mean would be 50g if the 'broken' scale actually works*
- *Treat the 10 packs as a sample (size 10) from the population*

As the population mean $\mu$ is concerned, a natural testing statistic is sample mean $\overline{X}$. Standardizing gives

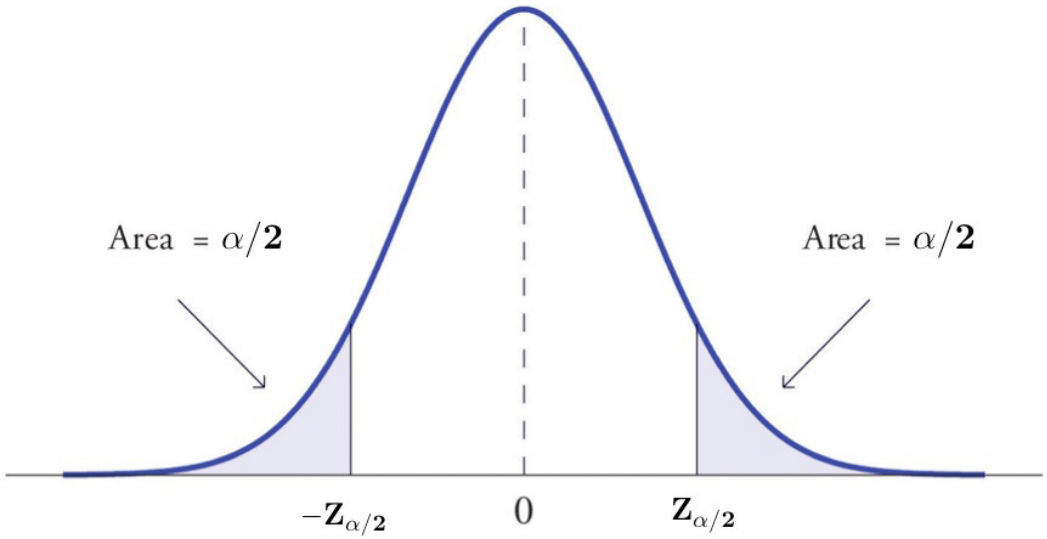
$$T = \frac{\overline{X} - \mu}{SD(\overline{X})} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

- $T \sim \mathcal{N}(0, 1)$ if $\sigma$ known and $n$ large (Z-test), otherwise
- $T \sim t_{n-1}$ (t-distribution with degrees of freedom $n - 1$, t-test).

# Example (continued)

Rejection region $\mathcal{R} = (-\infty, -Z_{\alpha/2}) \cup (Z_{\alpha/2}, \infty)$ (two-side test)

- Too heavy or too light of $\overline{X}$ would cause rejection of $H_0$
- Total size of the two tails is $\alpha$
- Lookup from t-table with d.f. 9 gives $Z_{\alpha/2} = 2.821$



# Example (continued)

t-table

Table entry for $p$ and $C$ is the critical value $t^*$ with probability $p$ lying to its right and probability $C$ lying between $-t^*$ and $t^*$.

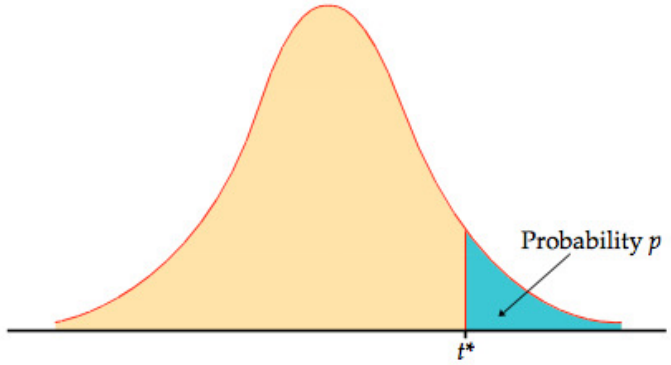



**TABLE D**

*t* distribution critical values

| | Upper-tail probability $p$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| df | .25 | .20 | .15 | .10 | .05 | .025 | .02 | .01 | .005 | .0025 | .001 | .0005 |
| 1 | 1.000 | 1.376 | 1.963 | 3.078 | 6.314 | 12.71 | 15.89 | 31.82 | 63.66 | 127.3 | 318.3 | 636.6 |
| 2 | 0.816 | 1.061 | 1.386 | 1.886 | 2.920 | 4.303 | 4.849 | 6.965 | 9.925 | 14.09 | 22.33 | 31.60 |
| 3 | 0.765 | 0.978 | 1.250 | 1.638 | 2.353 | 3.182 | 3.482 | 4.541 | 5.841 | 7.453 | 10.21 | 12.92 |
| 4 | 0.741 | 0.941 | 1.190 | 1.533 | 2.132 | 2.776 | 2.999 | 3.747 | 4.604 | 5.598 | 7.173 | 8.610 |
| 5 | 0.727 | 0.920 | 1.156 | 1.476 | 2.015 | 2.571 | 2.757 | 3.365 | 4.032 | 4.773 | 5.893 | 6.869 |
| 6 | 0.718 | 0.906 | 1.134 | 1.440 | 1.943 | 2.447 | 2.612 | 3.143 | 3.707 | 4.317 | 5.208 | 5.959 |
| 7 | 0.711 | 0.896 | 1.119 | 1.415 | 1.895 | 2.365 | 2.517 | 2.998 | 3.499 | 4.029 | 4.785 | 5.408 |
| 8 | 0.706 | 0.889 | 1.108 | 1.397 | 1.860 | 2.306 | 2.449 | 2.896 | 3.355 | 3.833 | 4.501 | 5.041 |
| 9 | 0.703 | 0.883 | 1.100 | 1.383 | 1.833 | 2.262 | 2.398 | 2.821 | 3.250 | 3.690 | 4.297 | 4.781 |
| 10 | 0.700 | 0.879 | 1.093 | 1.372 | 1.812 | 2.228 | 2.359 | 2.764 | 3.169 | 3.581 | 4.144 | 4.587 |
| 11 | 0.697 | 0.876 | 1.088 | 1.363 | 1.796 | 2.201 | 2.328 | 2.718 | 3.106 | 3.497 | 4.025 | 4.437 |
| 12 | 0.695 | 0.873 | 1.083 | 1.356 | 1.782 | 2.179 | 2.303 | 2.681 | 3.055 | 3.428 | 3.930 | 4.318 |
| 13 | 0.694 | 0.870 | 1.079 | 1.350 | 1.771 | 2.160 | 2.282 | 2.650 | 3.012 | 3.372 | 3.852 | 4.221 |
| 14 | 0.692 | 0.868 | 1.076 | 1.345 | 1.761 | 2.145 | 2.264 | 2.624 | 2.977 | 3.326 | 3.787 | 4.140 |
| 15 | 0.691 | 0.866 | 1.074 | 1.341 | 1.753 | 2.131 | 2.249 | 2.602 | 2.947 | 3.286 | 3.733 | 4.073 |
| 16 | 0.690 | 0.865 | 1.071 | 1.337 | 1.746 | 2.120 | 2.235 | 2.583 | 2.921 | 3.252 | 3.686 | 4.015 |

# Example (continued)

The value of the testing statistic is given by

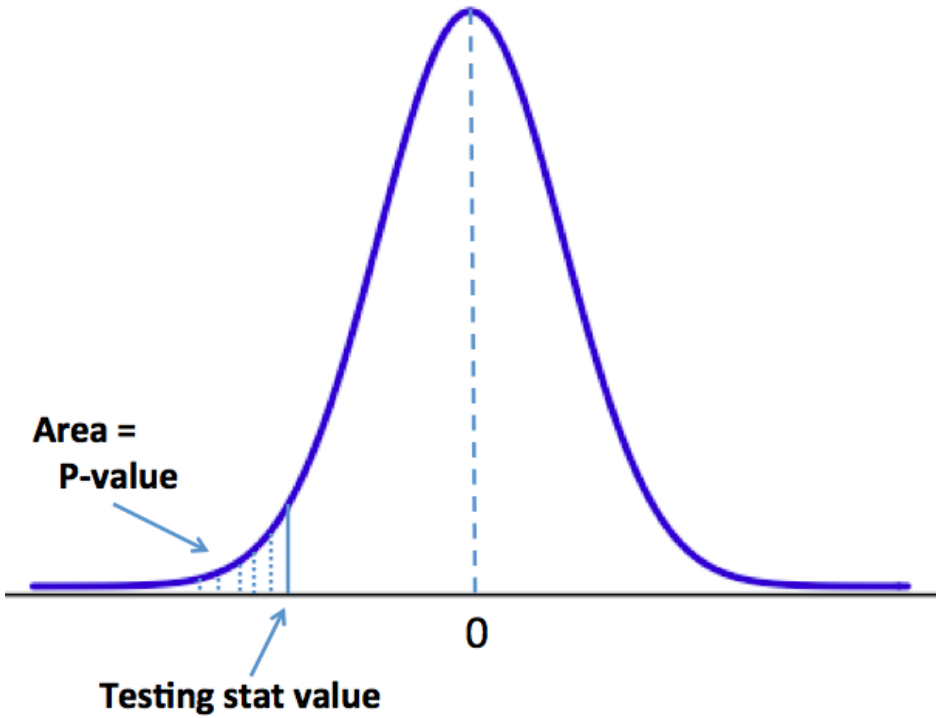
$$\frac{49.09 - 50}{s/\sqrt{n}} = \frac{-0.91}{0.7781/\sqrt{10}} = -3.6984 \in \mathcal{R}$$

where $\sigma$ is replaced by $s = \sqrt{\sum(X_i - \overline{X})^2/n}$

- So the conclusion is *reject* $H_0$ at $\alpha = 0.02$
  - *Interpretation*: at $\alpha = 0.02$, significant evidence suggesting that the 'broken' scale is indeed broken.
- Alternatively, one can calculate *p-value*
  - T-statistic of -3.6984 for df = 9 has a p-value of 0.004933
  - Reject $H_0$ as p-value $< \alpha$
  - Same conclusion as before (expected).

# P-value

- Often called *observed* p-value
  - Measures the significance of evidence against $H_0$
    - Smaller values indicate more significance
  - How likely to have testing stat value from sample as *extreme* as the calculated one.



# Example (one-side Z-test)

A street is considered *busy* if there are more than (included) 20 cars passing by within a non-peak hour. To determine if *Old Westport Road* in Dartmouth, MA is busy or not, Data Scientist Alex randomly picked 36 different non-peak hours and counted the number of cars passing by. The counts are

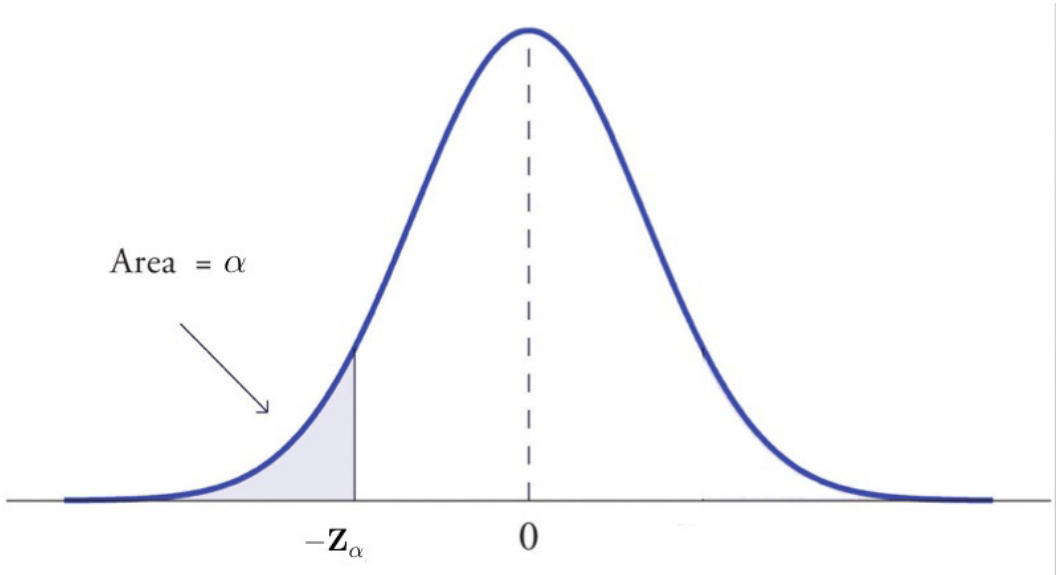
> *19, 20, 22, 24, 19, 19, 22, 25, 26, 20, 21, 24, 17, 15, 22, 25, 17, 18, 21, 20, 22, 18, 22, 21, 19, 20, 23, 21, 18, 17, 19, 21, 24, 21, 23, 20.*

Is the Old Westport Road a busy road?

# Example (continued)

Assume $\alpha = 0.05$. Null hypothesis $H_0 : \mu = 20$, $H_A : \mu < 20$.

- Here $H_0$ interpreted as $\mu \geq 20$, or, "Old Westport Road is busy"
  - Interpretation of $H_0$ is determined by the alternative, $H_A$
  - In $H_0$, we set $\mu = 20$ to make subsequent calculation easy
  - Rejection for small value of testing statistic.



# Example (continued)

Testing statistic

$$T = \frac{\overline{X} - \mu}{SD(\overline{X})} = \frac{\overline{X} - \mu}{s/\sqrt{n}} \approx \mathcal{N}(0, 1)$$

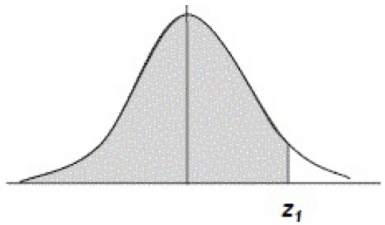
- Here t- and Z-table very similar as sample size large enough

Rejection region is $\mathcal{R} = (-\infty, -1.64)$, and testing statistic

$$\frac{20.6944 - 20}{s/\sqrt{n}} = \frac{0.6944}{2.5615/\sqrt{36}} = 1.63 \notin \mathcal{R}.$$

So we *don't reject* null hypothesis $H_0$ at $\alpha = 0.05$.

# Standard normal table




$$p(z \le z_1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_1} e^{-\frac{1}{2}z^2} dz$$

| $z_1$ | 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 |
| 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 |
| 2.1 | 0.9821 | 0.9826 | 0.9830 | 0.9834 | 0.9838 | 0.9842 | 0.9846 | 0.9850 | 0.9854 | 0.9857 |
| 2.2 | 0.9861 | 0.9864 | 0.9868 | 0.9871 | 0.9875 | 0.9878 | 0.9881 | 0.9884 | 0.9887 | 0.9890 |
| 2.3 | 0.9893 | 0.9896 | 0.9898 | 0.9901 | 0.9904 | 0.9906 | 0.9909 | 0.9911 | 0.9913 | 0.9916 |
| 2.4 | 0.9918 | 0.9920 | 0.9922 | 0.9925 | 0.9927 | 0.9929 | 0.9931 | 0.9932 | 0.9934 | 0.9936 |

# Type I error

- In hypothesis testing framework, when one rejects $H_0$
  - This could mean that sth wrong with $H_0$ (so reject it)
  - Any other possibility?

# Type I error (continued)

- In hypothesis testing framework, when one rejects $H_0$
  - This could mean that sth wrong with $H_0$
    - So reject it
  - Or $H_0$ is true, but one is unlucky in getting *bad* data
    - Whenever this happens, one makes mistake in rejecting $H_0$
    - Term such error as *Type I error*, and

$$\mathbb{P}(\text{Type I error}) = \alpha.$$

# Type II error

- Another type of error occurs if one rejects $H_A$ when $H_A$ is true
  - Probability of Type II error is denoted by $\beta$
  - $1 - \beta$ is called the power of a hypothesis testing
  - It doesn't make sense to write $\alpha + \beta$.

| | When $H_0$ is true | When $H_A$ is true |
|---|---|---|
| Reject $H_0$ | *Type I error* | Correct inference |
| Reject $H_A$ | Correct inference | *Type II error* |

# An analogy to court trial

| | **Hypothesis testing** | **Court trial** |
|---|---|---|
| Assumption | $H_0$ is true | The defendant is not guilty |
| Action | Collect data sample | Collect evidence |
| Expect | Small chance event<br>(getting extreme values) | Surprise<br>(evidence against the defendant) |
| Decision | Rejection of $H_0$ | Rejection of assumption |
| Correctness | Possibility of chance error | Possibility of error |

# Type I and II error in context of court trials

| | When $H_0$ is true<br>*Defendant is not guilty* | When $H_A$ is true<br>*Defendant is indeed guilty* |
|---|---|---|
| Reject $H_0$ | Type I error<br>*Declare defendant guilty* | Correct decision<br>*Declare defendant guilty* |
| Reject $H_A$ | Correct decision<br>*Declare defendant not guilty* | Type II error<br>*Declare defendant not guilty* |

# 2-sample test

Discussion so far on hypothesis testing

- All deals with parameters of one population
  - e.g., population mean $\mu$
  - A sample from the population required for inference
  - One-sample test
- There are situations two different populations involved
  - e.g., A/B test in model assessment
  - A sample from each population is required
  - ♠ *Two-sample test.*

# Reduction of proportion to mean

- Introduce r.v. as follows

$$X_i = \begin{cases} 1, & \text{success of event i} \\ 0, & \text{otherwise} \end{cases}$$

  - This is a *Bernouli* random variable, assume $P(X_i = 1) = p$
  - Outcome of one trial

- Thus $X_1 + X_2 + ... + X_n =$ total # success out of n trials
  - Proportion of success expressed as

$$T = (X_1 + X_2 + ... + X_n)/n$$

  - This is the familiar sample mean
    - $\mathbb{E}(T) = p,$ and $SD(T) = SD(X_1)/n = p(1-p)/n.$

# An example on 2-sample test

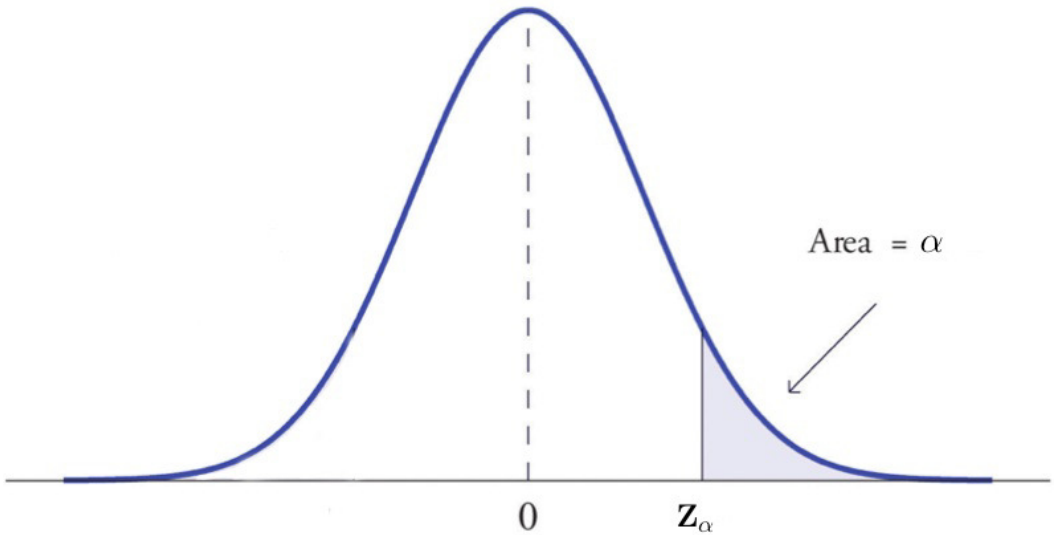
Data scientist *Alex* recently developed a new model. To determine if the model is better than current one, the quality team did the so-called *A/B test*. They randomly split the user access to the two model versions. Out of the 200 user access under the new model, there are 96 clicks of products; the 600 access to the current model leads to 240 clicks of products.

- *Is the new model better? Assume* $\alpha = 0.05$.

# Example (continued)

Null hypothesis $H_0 : \mu_n - \mu_c = 0$, $H_A : \mu_n - \mu_c > 0$.

- Here $H_0$ interpreted as $\mu_n - \mu_c \leq 0$, i.e., current model better
- Large value of $\overline{X} - \overline{Y}$ would cause rejection of $H_0$.



# Example (continued)

Testing statistic

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_n - \mu_c)}{SD(\overline{X} - \overline{Y})} \sim t_{798}$$

Rejection region is $\mathcal{R} = (1.64, \infty)$, and testing statistic

$$\frac{0.48 - 0.40}{\sqrt{0.48 * 0.52/n_1 + 0.4 * 0.6/n_2}} = 1.9707 \in \mathcal{R}.$$

So we *reject* $H_0$ at $\alpha = 0.05$, or, the new model is better.

# R function t.test()

- Can be used for both one and two-sample t-test
- The broken scale example

```
> t.test(x,alternative=c("two.sided","less","greater"), mu=50)

One Sample t-test

data:  x
t = -3.6983, df = 9, p-value = 0.004933
alternative hypothesis: true mean is not equal to 50
95 percent confidence interval:
 48.53338 49.64662
sample estimates:
mean of x
    49.09
```

# R function t.test() (continued)

- The busy street example

```
> t.test(x,alternative="less",mu=20)

One Sample t-test

data:  x
t = 1.6267, df = 35, p-value = 0.9436
alternative hypothesis: true mean is less than 20
95 percent confidence interval:
     -Inf 21.41574
sample estimates:
mean of x
 20.69444
```

# R function t.test() (continued)

- The A/B test example

```
> x<-c(rep(1,96),rep(0,200-96))
> y<-c(rep(1,240),rep(0,600-240))
> t.test(x,y,alternative="greater",var.equal=TRUE)

Two Sample t-test

data:  x and y
t = 1.9876, df = 798, p-value = 0.0236
alternative: true difference in means is greater than 0
95 percent confidence interval:
 0.01371795        Inf
sample estimates:
mean of x mean of y
     0.48      0.40
```

# Practice I

The PM2.5 is an important index for air pollution. A PM2.5 index of over 100 is harmful. To decide if the air over an industry zone is harmful, one took PM2.5 readings at 35 randomly picked spots in the zone. The readings are

*119, 120, 102, 94, 79, 119, 92, 85, 96, 120, 101, 84, 97, 115, 107, 125, 117, 98, 91, 80, 122, 108, 122, 111, 109, 90, 83, 101, 118, 107, 99, 80, 102, 97, 96.*

Is the air in the industry zone harmful? Assume $\alpha = 0.02$.

# Practice II

You are given a coin. You do not know if it is a fair coin or not. You want to figure out a way so that you can make a sound statement about the fairness of the coin.