population and sample

- Objective: We use statistics is to estimate the characteristics of the population

- Data is often reported as **X ± ΔX.** How do we determine **X** and **ΔX** ?
- What are the meanings of uncertainty?
- How do they combine through additive and multiplicative operations?

1

Measurements and uncertainty

- One of the best ways to understand the characteristics of the uncertainty is to measure several times and evaluate the differences in measurements

- **Systematic error:** this type of error cannot be treated with statistics. Calibration with known standard is required to reduce the level of systematic error to an acceptable level.

- **Random error:** random uncertainty is revealed by repeat sampling, and can be treated with statistics. Most of today's discussion deals with this type of error.

2

What is the representative value?

- Mean
- Median (50 percentile)

3

How much does the data vary?

- Standard deviation
- IQR (Inter-Quartile Range) = (75 percentile) – (25 percentile)

4

Mean vs median

- Mean and Median: which one should you choose?

- Atlanta mean August temperature
- **Mean =** 78.335211126760563 °F
- **Median = 78.2** °F

Significant figure should reflect the precision of the measurements

	Year	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
0	1879	44.3	43.7	57.6	58.9	69.8	74.9	79.6	73.9	68.7	64.3	53.9	51.5
1	1880	54.3	51.4	55.5	63.4	71.0	76.5	79.2	76.8	69.4	60.8	47.2	42.1
2	1881	40.1	46.6	49.1	59.1	70.8	77.8	81.1	78.8	75.5	67.2	52.6	49.6
3	1882	48.8	52.4	57.4	64.8	66.1	76.6	75.9	75.8	71.7	65.8	51.1	41.4

5

Mean vs median

- Mean and Median: which one should you choose?

- 9 people makes $50K and 1 person make $1M
- Mean = $145K
- Median = $50K

- The answer depends on the statistical distribution of the data. Mean and variance are impacted by outliers.

6

Mean vs median

- Mean and Median: which one should you choose?

- 9 people makes $50K and 1 person make $1M
- Remove outlier ($1M) and re-calculate the mean and median

- Mean (w/o outlier): $50K
- Median: $50K
  - Data quality control (flagging outliers) may be necessary before taking the mean and standard deviation

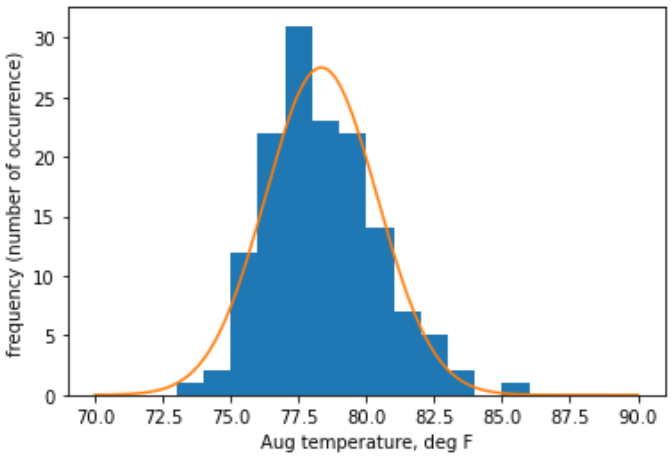
7

Statistical distribution

- Statistical distribution of data: Histogram

Orange line is the Gaussian, g(x), multiplied by the sample size

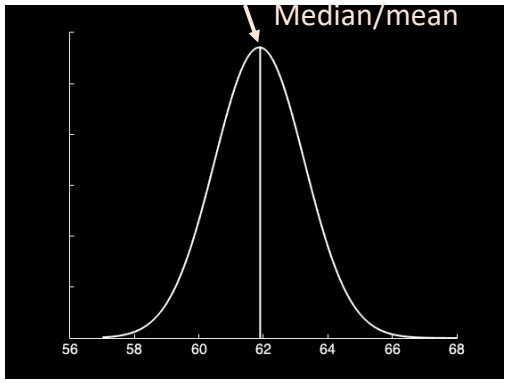
$$g(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$



8

Gaussian (normal distribution)

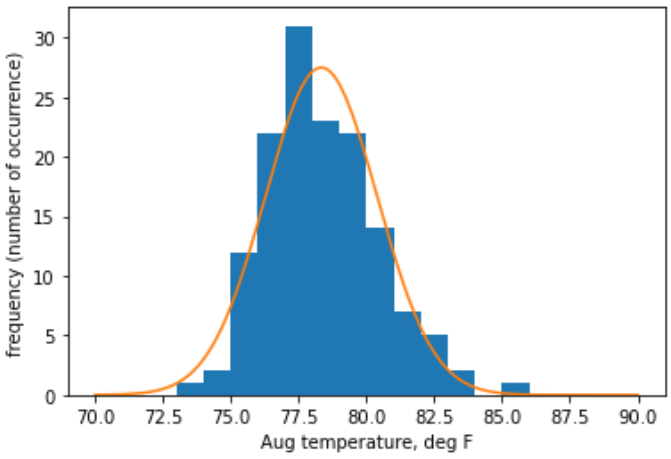
- Atlanta's monthly mean temperature ≃ Gaussian
- Properties of Gaussian (normal distribution)
- Symmetric
- Mean = Median
- $\mu \pm 1\sigma$ contains 68% of data
- $\mu \pm 2\sigma$ contains 95% of data



9

Statistical distribution

- **Uncertainty of single measurement:**
- IF the data is normally distributed, there is 95% probability that a single measurement (x) lies within the range of 2 standard deviation (σ) from the population mean (μ).
- $x = \mu \pm 2\sigma$



10

f(x) = PDF = Probability Density Function

- PDF is a function that describes statistical distribution
- Its area (integral) measures the probability

$$P(x < b) = \int_{-\infty}^{b} f(x)\,dx$$

- Graph of f(x) is different from the histogram that PDF integrates out to **1**. The area covered by the histogram is equal to **NΔx** where **N** is the sample size and **Δx** is the bin size.

11

Statistical distribution

- What about the standard deviation?

$$\sigma^2 = \frac{(1-3.5)^2 + (2-3.5)^2 + \cdots + (6-3.5)^2}{6} \sim 2.9$$

$$= \Sigma(x-\mu)^2 P(x)$$

$$\sigma \sim 1.7$$

12

Central Limit Theorem

- 1. Mean of sample mean (M) is equal to the population mean (μ)
- 2. Standard error $\sigma/\sqrt{N}$ measures the average distance between sample mean (M) and the true mean (μ)
- 3. Sample mean (M) follows the normal distribution (Gaussian) regardless of the parent population.

13

Uncertainty from multiple (N) measurements

- Consider taking **N samples randomly from a population**
- What is the 95% confidence interval on the **true mean**?

- Sample mean (M) is the average of N samples
- Uncertainty scales with SE. $SE = \frac{s}{\sqrt{N-1}}$
- If N > 30, 95% confidence interval is $\mu = M \pm 2SE$

14

Uncertainty from multiple (N) measurements

- Consider taking **N samples randomly from a population**
- What is the 95% confidence interval on the **true mean**?

- Sample mean (M) is the average of N samples
- Uncertainty scales with SE. $SE = \frac{s}{\sqrt{N-1}}$
- If N < 30, use Student's t-distribution instead of Gaussian.
- The 95% confidence interval is $\mu = M \pm t_{crit} SE$

15

Propagation of uncertainty

- Additive problem
- **D = X + 2Y – 3Z**
- Given the best estimates and uncertainties in (X, Y, Z), estimate the range of D
- The most conservative uncertainty estimate:

$$\delta D < \delta X + 2\delta Y + 3\delta Z$$

16

Propagation of uncertainty

- Additive problem
- **D = X + 2Y – 3Z**
- Given the best estimates and uncertainties in (X, Y, Z), estimate the range of D
- If the uncertainty is independent and random:

$$\delta D = \sqrt{(\delta X)^2 + (2\delta Y)^2 + (3\delta Z)^2}$$

17

Propagation of uncertainty

- Multiplicative problem
- **D = X * Y / Z**
- Given the best estimates and uncertainties in (X, Y, Z), estimate the range of D
- The most conservative uncertainty estimate:

$$\frac{\delta D}{|D|} < \frac{\delta X}{|X|} + \frac{\delta Y}{|Y|} + \frac{\delta Z}{|Z|}$$

18

Propagation of uncertainty

- Multiplicative problem
- **D = X * Y / Z**
- Given the best estimates and uncertainties in (X, Y, Z), estimate the range of D
- If the uncertainty is independent and random:

$$\frac{\delta D}{D} = \sqrt{\left(\frac{\delta X}{X}\right)^2 + \left(\frac{\delta Y}{Y}\right)^2 + \left(\frac{\delta Z}{Z}\right)^2}$$

19

Hypothesis testing

The 5 steps

1. State your confidence level
2. State your null hypothesis and its alternative
3. State the statistics used
4. Determine the critical region
5. Evaluate whether the data is within or outside of the critical region

20

1. State your confidence level

- It is important to set your confidence level first. This sets how extreme the data should be in order to reject the null hypothesis.

Example

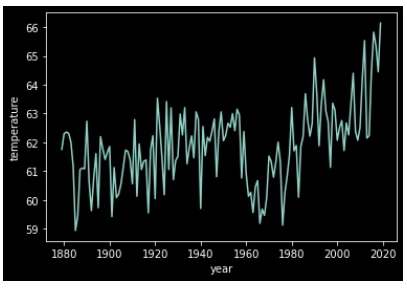
- 95% confidence level means we reject null hypothesis if data is extreme beyond 95% of expected range.

21

2. State the null hypothesis and its alternative

Example

- H_0: Annual mean temperature from the last 10 years is NOT significantly warmer than the long-term mean.
- H_1: Annual mean temperature from the last 10 years is significantly warmer than the long-term mean.



22

3. State the statistics used.

Example

- We use Student's t-distribution with one-tail test.
- Student's t-distribution: In general when the sample number is small (N<30) we use t-distribution.
- One-tail test: Since we already know that some warming has occurred, we focus on the positive side of the distribution.

23

4. Determine the critical region

- Critical region is related to the percentile value.
- For one-tail test, the critical region with the 95% confidence level is 95% percentile value
- Look for the t-value at 0.95 = CDF(t,d.f.=9) where CDF is the cumulative distribution function.
- This can be done using MATLAB/python

```
>> a = 1-.95
>> tcrit = tinv(1-a,df)        (for one tail)
>> tcrit = tinv(1-a/2,df)      (for two tail)
```

24

5. Evaluate whether data is within the critical region

- Calculate the t-value of the data:
- x: Mean temperature from the last 10 years
- μ: Long term mean temperature
- t-value is (signal)/(standard error)

$$\text{t-value} = \frac{x - \mu}{s/\sqrt{N-1}}$$

25

5. Evaluate whether data is within the critical region

- If the null hypothesis is correct, the t-value will be $t<t_{crit}$ for the confidence level (95%)
- $t>t_{crit}$: we reject the null hypothesis
- Random sampling cannot explain the warm temperature from the last 10 years!
- (p-value) What's the probability of getting the observed signal by random sampling: p=1-CDF(tvalue,df)
  p=0.0005 (0.05%).

26

Type I vs II errors

- There are 4 possible outcomes of the hypothesis testing
- 1. Null hypothesis is NOT true, and the t-test rejects the null hypothesis (true positive)
- 2. Null hypothesis is true, and the t-test does not reject the null hypothesis (true negative)
- 3. Null hypothesis is true, and the the t-test rejects the null hypothesis (false positive = Type I error)
- 4. Null hypothesis is NOT true, and the t-test does not reject the null hypothesis (false negative = Type II error)

27

Type I error

- False positive

- Random sampling can sometimes create a strong signal.
- If confidence level is 95%, this happens at 5% probability (the "α" value is the probability of committing Type I error)

- To reduce the Type I error, increase the confidence level (i.e. lower the α value). Higher level of confidence requires a stronger t-value in order to reject H_0.

28

Type II error

- False negative

- The signal is not very strong and/or sample size is small.
- If the confidence level is very high, type II error can occur.

- To reduce the Type II error while keeping certain confidence level, it is necessary to collect more data. You should not just lower the confidence level to get the result you want.

29

Comparing two sample statistics

- Assume you want to compare 2 groups of samples

- Sample size N_1 and N_2, means x_1 and x_2, standard deviation s_1 and s_2
- Null hypothesis: the two samples are coming from the same population ($x_1 = x_2$)

$$t = \frac{x_1 - x_2}{\sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \qquad \sigma = \sqrt{\frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}}$$

30