Basic Statistics

Contents

1.1 Mean, variance and higher moments

You have a random variable, let's call it X, and you draw individual values of X, denoted $[x_1, x_2, x_3, ...x_N]$.

We define

$$\mu = \text{the population mean} \tag{1}$$
$$\sigma = \text{the population standard deviation} \tag{2}$$
$$\overline{x} = \text{the sample mean} \tag{3}$$
$$s = \text{the same standard deviation} \tag{4}$$

1.1.1 Mean

The sample mean is defined as

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{5}$$

where the bar denotes the time mean and the subscript i denotes the time step (we will be working mainly with time series here).

The sample mean $\overline{x}$ is an unbiased estimate of the true mean μ.

In other words: if you draw an infinite number of *samples* from the same time series, then the actual population mean of all of the sample means ($\mu_{\overline{x}}$) is equal to the population mean (μ).

The mean is the first moment about zero.

1.1.2 Median

The median is the value in the center of the population. Useful when you have large outliers in the distribution (e.g. house prices, salaries, rain rate).

1.1.3 Variance

The sample variance is defined as:

$$\overline{x'^2} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2 \tag{6}$$

where the prime denotes departures from the mean.

Example: VARIANCE DEFINITION

The sample standard deviation is defined as:

$$s = \sqrt{\overline{x'^2}} \tag{7}$$

The variance is the second moment about the mean.

The division by N-1 is required to obtain an unbiased estimate of the true population variance.

The bias creeps in because we have to estimate the sample mean to get the sample variance, and so we lose one of our degrees of freedom (more on this later). Another way to see this is:

$$\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2 = \frac{1}{N}\sum_{i=1}^{N}((x_i - \mu) - (\overline{x} - \mu))^2 \tag{8}$$

What we want is the first term, and the second term is our bias. You can show that adding the $N-1$ gets rid of the low bias of the variance, but I don't want to go into that derivation.

In practice, if using N or N-1 matters in your analysis, you are in trouble anyway.

The mean is the average picture (stationary picture), while the variance describes some of the interesting wiggles and variability about the mean.

The estimate of the variance depends greatly on the sample rate (e.g. if you are using daily data, monthly data, yearly data, etc).

Example: PYTHON NOTEBOOK VARIANCE_EXAMPLE.IPYNB

Also, one can show that two time series with the same variance (say, of 1) but with different persistences can give different answers depending on the sample window (how long of the record you look).

1.1.4 Higher moments

In general, all moments are defined as:

$$m_r = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^r \tag{9}$$

m_2 is the variance; m_3 is the skewness; m_4 is the kurtosis;

These can also be given in non dimensional form as:

$$a_r = m_r / s^r \tag{10}$$

where s is the sample standard deviation.

The skewness indicates the degree of asymmetry of the distribution about the mean. $a_3 > 0$ indicates a long tail on the positive side of the mean (a gaussian has $a_3 = 0$).

Example: SKEWNESS SKETCH

The kurtosis indicates the degree of asymmetry of the distribution about the mean. $\mathfrak{a}_4 > 3$ indicates a more peaked distribution compared to a gaussian, which has $\mathfrak{a}_4 = 3$.

Example: KURTOSIS SKETCH

Typically, higher order moments are given in terms of the standardized units, for ease of comparison (and so units don't matter).

1.2 Basic probabilities; unions; intersections; conditional probabilities; Bayes theorem

1.2.1 Probabilities: Unions & Intersections

Define some event, E, for example, E is that you role a die and get a 2.

For a fair die, we write that the probability of E,

$$\boldsymbol{Pr}(\mathsf{E}) = \frac{1}{6} \tag{11}$$

The probability of not E happening, or,

$$\boldsymbol{Pr}\left(\tilde{\mathsf{E}}\right) = 1 - \boldsymbol{Pr}(\mathsf{E}) \tag{12}$$

For the case of rolling the die,

$$\boldsymbol{Pr}\left(\tilde{\mathsf{E}}\right) = 1 - \frac{1}{6} = \frac{5}{6} \tag{13}$$

The probability that two events, E_1 (rolling a 5) and E_2 (rolling a 6) will *both* occur is called the intersection of the two probabilities and is written

$$\boldsymbol{Pr}(\mathsf{E}_1 \cap \mathsf{E}_2) \tag{14}$$

The probability that either or both of two events, E_1 (rolling a 5) and E_2 (rolling a 6) will occur is called the union of the two probabilities (not to be confused with XOR, which is an *exclusive OR*, meaning only one can be true). The union is written

$$\boldsymbol{Pr}(\mathsf{E}_1 \cup \mathsf{E}_2) = \boldsymbol{Pr}(\mathsf{E}_1) + \boldsymbol{Pr}(\mathsf{E}_2) - \boldsymbol{Pr}(\mathsf{E}_1 \cap \mathsf{E}_2) \tag{15}$$

One can see why the intersection term comes into the union equation by drawing a Venn Diagram

Example: VENN DIAGRAM

In this diagram the area in the rectangle represents the total probability of one, and the area inside the two event circles indicates the probability of the two events. The intersection

between them gets counted twice when you add the two areas and so must be subtracted to calculate the union of the probabilities. If the two events are mutually exclusive, then no intersection occurs, i.e., if one happens, we know the other did not happen at that time.

Another important concept is conditional probability,

$$\boldsymbol{Pr}(\mathsf{E}_2|\mathsf{E}_1) = \frac{\boldsymbol{Pr}(\mathsf{E}_1 \cap \mathsf{E}_2)}{\boldsymbol{Pr}(\mathsf{E}_1)} \tag{16}$$

read "E_2 given E_1".

Changing this conditional probability relationship around a little yields a formula for the probability that both events will occur, called the multiplicative law of probability:

$$\boldsymbol{Pr}(\mathsf{E}_1 \cap \mathsf{E}_2) = \boldsymbol{Pr}(\mathsf{E}_2|\mathsf{E}_1) \cdot \boldsymbol{Pr}(\mathsf{E}_1) = \boldsymbol{Pr}(\mathsf{E}_1|\mathsf{E}_2) \cdot \boldsymbol{Pr}(\mathsf{E}_2) \tag{17}$$

If E_1 and E_2 are independent events (their probabilities don't depend on one another, like rolling a fair die), then

$$\boldsymbol{Pr}(\mathsf{E}_2|\mathsf{E}_1) = \boldsymbol{Pr}(\mathsf{E}_2) \tag{18}$$

and thus

$$\boldsymbol{Pr}(\mathsf{E}_1 \cap \mathsf{E}_2) = \boldsymbol{Pr}(\mathsf{E}_2) \cdot \boldsymbol{Pr}(\mathsf{E}_1) \tag{19}$$

This is the definition of statistically independent.

Rolling a fair die yields statistically independent events, and thus, the probability of rolling both a 5 and a 6 is equal to $\frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$.

Note, if the two events are not independent, multiplying the probabilities will give the wrong answer!

1.2.1.1 Example of conditional probability

Say the probability of it raining on Monday is 60%. But, you know from looking at historical records that the probability of it raining the day after it rains is 80% (it is more likely than not to rain the day after it rains). So, whether it rains on Tuesday is dependent on whether it rains on Monday. What is the probability it will rain Monday and Tuesday?

$$\mathsf{M} = \text{event: it rains Monday} \tag{20}$$

$$\mathsf{T} = \text{event: it rains Tuesday} \tag{21}$$

$$\boldsymbol{Pr}(\mathsf{M} \cap \mathsf{T}) = \boldsymbol{Pr}(\mathsf{T}|\mathsf{M}) \cdot \boldsymbol{Pr}(\mathsf{M}) = 0.8 \cdot 0.6 = 48\% \tag{22}$$

1.2.2 Bayes Theorem

Let $\mathsf{E}_\mathsf{i}, \mathsf{i} = 1, 2, 3...\mathsf{N}$ be a set of N events such that the set E_i includes all possible possibilities in a set S and the events are mutually exclusive. Then, for any event B, with $\boldsymbol{Pr}(\mathsf{B}) > 0$

$$\boldsymbol{Pr}(\mathsf{E}_\mathsf{j}|\mathsf{B}) = \frac{\boldsymbol{Pr}(\mathsf{B}|\mathsf{E}_\mathsf{j})\,\boldsymbol{Pr}(\mathsf{E}_\mathsf{j})}{\sum_{\mathsf{i}=1}^{\mathsf{N}} \boldsymbol{Pr}(\mathsf{B}|\mathsf{E}_\mathsf{i})\,\boldsymbol{Pr}(\mathsf{E}_\mathsf{i})} \tag{23}$$

Note that since the E_is cover all possible outcomes, the following is true:

$$\boldsymbol{Pr}(\mathsf{B}) = \sum_{\mathsf{i}=1}^{\mathsf{N}} \boldsymbol{Pr}(\mathsf{B}|\mathsf{E}_\mathsf{i})\,\boldsymbol{Pr}(\mathsf{E}_\mathsf{i}) \tag{24}$$

Note that this is the denominator in Bayes theorem.

A useful relation when doing Bayes Theorem problems is the following:

$$\boldsymbol{Pr}(\mathsf{E}_2|\mathsf{E}_1) = 1 - \boldsymbol{Pr}\left(\tilde{\mathsf{E}}_2|\mathsf{E}_1\right) \tag{25}$$

In general, Bayes' Theorem takes $\boldsymbol{Pr}(A|\mathsf{B})$ and turns it into $\boldsymbol{Pr}(\mathsf{B}|A)$.

Let's do an example.

1.2.2.1 Example of Bayes Theorem: testing for a chemical

You have designed a test which tells whether a sample of air has a harmful airborne chemical or not. The probability that any individual sample has the chemical is 1/1000. The test never gives a false negative result (saying that the sample does not have the chemical, when in fact it does). The chance of a false positive is also small, say 5%. If you test a sample of air today and the test comes back positive, what is the probability that the chemical was actually in the sample?

Your instincts might tell you 95%, since the false positive rate is 5%. However, this is not correct, since you need to take into account all of the information. To do these problems, it is good to follow these steps:

1. Define your variables
2. Clearly state what you want to know
3. List all of the information the problem gives you
4. Check the assumptions for your method of solving
5. Then, solve for what you listed in Step 2

(1) Define the variables:

$$\boldsymbol{Pr}(\mathsf{C}) = \text{probability the chemical is present} \tag{26}$$
$$\boldsymbol{Pr}\left(\tilde{\mathsf{C}}\right) = \text{probability the chemical is not present} \tag{27}$$
$$\boldsymbol{Pr}(+) = \text{probability the test is positive} \tag{28}$$
$$\boldsymbol{Pr}\left(\tilde{+}\right) = \text{probability the test is negative} \tag{29}$$

(2) State what we want to know:

$$\boldsymbol{Pr}(\mathsf{C}|+) \tag{30}$$

(3) List what we know

$$\boldsymbol{Pr}(\mathsf{C}) = 0.001 \tag{31}$$
$$\boldsymbol{Pr}\left(\tilde{+}|\mathsf{C}\right) = 0 \tag{32}$$
$$\boldsymbol{Pr}\left(+|\tilde{\mathsf{C}}\right) = 0.05 \tag{33}$$

(4) Check assumptions. Here, the set S is made of two possibilities that span all possibilities,

$$\mathsf{E}_1 = \mathsf{C} = \text{the chemical is present} \tag{34}$$
$$\mathsf{E}_2 = \tilde{\mathsf{C}} = \text{the chemical is not present} \tag{35}$$

(5) We can now apply Bayes Theorem.

$$\boldsymbol{Pr}(\mathsf{C}|+) = \frac{\boldsymbol{Pr}(+|\mathsf{C})\,\boldsymbol{Pr}(\mathsf{C})}{\boldsymbol{Pr}(+|\mathsf{C})\,\boldsymbol{Pr}(\mathsf{C}) + \boldsymbol{Pr}\left(+|\tilde{\mathsf{C}}\right)\boldsymbol{Pr}\left(\tilde{\mathsf{C}}\right)} \tag{36}$$

We know all values, except, $\boldsymbol{Pr}(+|\mathsf{C})$. However, this is equal to

$$1 - \boldsymbol{Pr}\left(\tilde{+}|\mathsf{C}\right) = 1 - 0 = 1 \tag{37}$$

Plugging in the values leads to $\boldsymbol{Pr}(\mathsf{C}|+) = \frac{1}{51} \approx 2\%$. This result shows how important it is to take into account the background rate - otherwise, you may have thought the probability was 95%, not 2%.

1.2.2.2 Example of Bayes Theorem: cab accidents

The next problem comes from Chapter 16 of "Thinking Fast and Slow" by Daniel Kahneman, a popular science book on how our brains process information.

A cab was involved in a hit and run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data:

- 85% of the cabs in the city are Green and 15% are Blue.
- A witness identified the cab as Blue. The court tested the reliability of the witness under the circumstances that existed the night of the accident and concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time.

What is the probability that the cab involved in the accident was Blue rather than Green?

Your instincts might tell you 80% - but then you wouldn't be taking into account the background rate of the cabs in the city.

Note that if there hadn't been a witness, your answer would have been 15% (this is a frequentist approach - more on this later).

(1) Define the variables:

$$\boldsymbol{Pr}(\mathsf{B}) = \text{probability the cab was Blue} \tag{38}$$

$$\boldsymbol{Pr}\left(\tilde{\mathsf{B}}\right) = \text{probability the cab was not Blue (i.e. the cab was Green)} \tag{39}$$

$$\boldsymbol{Pr}(W) = \text{probability the witness witnessed a Blue cab} \tag{40}$$

(2) State what we want to know:

$$\boldsymbol{Pr}(\mathsf{B}|W) \tag{41}$$

(3) List what we know:

$$\boldsymbol{Pr}(\mathsf{B}) = 0.15 \tag{42}$$

$$\boldsymbol{Pr}\left(\tilde{\mathsf{B}}\right) = 0.85 \tag{43}$$

$$\boldsymbol{Pr}(W|\mathsf{B}) = 0.8 \tag{44}$$

$$\boldsymbol{Pr}\left(W|\tilde{\mathsf{B}}\right) = 0.2 \tag{45}$$

(4) Check assumptions:
OK.
(5) Solve for what you listed in Step (2):
Apply Bayes Theorem for $\boldsymbol{Pr}(\mathsf{B}|W)$.

$$\boldsymbol{Pr}(\mathsf{B}|W) = \frac{\boldsymbol{Pr}(W|\mathsf{B})\,\boldsymbol{Pr}(\mathsf{B})}{\boldsymbol{Pr}(W|\mathsf{B})\,\boldsymbol{Pr}(\mathsf{B}) + \boldsymbol{Pr}\left(W|\tilde{\mathsf{B}}\right)\boldsymbol{Pr}\left(\tilde{\mathsf{B}}\right)} \tag{46}$$

$$\boldsymbol{Pr}(\mathsf{B}|W) = \frac{0.8 \cdot 0.15}{0.8 \cdot 0.15 + 0.2 \cdot 0.85} = 0.41 \tag{47}$$

1.2.2.3 Example of Bayes Theorem: Monty Hall problem

This is a famous one, called the Monty Hall Problem and loosely based on a gameshow called "Let's Make A Deal" with host Monty Hall.

Suppose you're on a game show, and you're given the choice of three doors (A,B,C): Behind one door is a car; behind the others, goats (and you don't like goats). You pick a door, say door A, and the host, who knows what's behind the doors, opens another door, say door B, which has a goat.

Monty then says to you, "Do you want to switch your choice and pick door C?" Is it to your advantage to switch your choice?

To start, imagine there is no host, and you have to choose the door with the car. You choose door A, what is the probability that you are correct? Easy - $\frac{1}{3}$.

Monty showing you what is behind one of the doors is extra information. This should tell you, perhaps I should use Bayes Theorem! (Drawing pictures can help too). The hardest part of this problem is figuring out how the set it up.)
Without loss of generality, we will assume that you have chosen door A.

(1) Define the variables:

$$\begin{aligned}
\boldsymbol{Pr}(\mathsf{A}) &= \text{probability the car is behind door A} && (48)\\
\boldsymbol{Pr}(\mathsf{B}) &= \text{probability the car is behind door B} && (49)\\
\boldsymbol{Pr}(\mathsf{C}) &= \text{probability the car is behind door C} && (50)\\
\boldsymbol{Pr}(\mathsf{M}) &= \text{probability Monty Hall opens door B} && (51)\\
& && (52)
\end{aligned}$$

(2) State what we want to know:

$$\boldsymbol{Pr}(\mathsf{C}|\mathsf{M}) \tag{53}$$

(3) List what we know:

$$\begin{aligned}
\boldsymbol{Pr}(\mathsf{A}) &= \boldsymbol{Pr}(\mathsf{B}) = \boldsymbol{Pr}(\mathsf{C}) = \frac{1}{3} && (54)\\
\boldsymbol{Pr}(\mathsf{M}|\mathsf{A}) &= \frac{1}{2} \text{ he could open B or C, since the prize is behind A} && (55)\\
\boldsymbol{Pr}(\mathsf{M}|\mathsf{B}) &= 0 && (56)\\
\boldsymbol{Pr}(\mathsf{M}|\mathsf{C}) &= 1 && (57)\\
&\rightarrow \text{ You chose door A} && (58)
\end{aligned}$$

(4) Check assumptions. OK.
(5) We can now apply Bayes Theorem.

$$\boldsymbol{Pr}(\mathsf{C}|\mathsf{M}) = \frac{\boldsymbol{Pr}(\mathsf{M}|\mathsf{C})\,\boldsymbol{Pr}(\mathsf{C})}{\boldsymbol{Pr}(\mathsf{M}|\mathsf{C})\,\boldsymbol{Pr}(\mathsf{C}) + \boldsymbol{Pr}(\mathsf{M}|\mathsf{B})\,\boldsymbol{Pr}(\mathsf{B}) + \boldsymbol{Pr}(\mathsf{M}|\mathsf{A})\,\boldsymbol{Pr}(\mathsf{A})} \tag{59}$$

$$\boldsymbol{Pr}(\mathsf{C}|\mathsf{M}) = \frac{1 \cdot 1/3}{1 \cdot 1/3 + 0 \cdot 1/3 + 1/2 \cdot 1/3} = \frac{1/3}{1/3 + 1/6} = \frac{1/3}{1/2} = \frac{2}{3} \tag{60}$$

$$\tag{61}$$

So, if you switch your choice, you have a higher probability of winning the car!

1.2.3 Probability philosophy

There are two general philosophies of thought on statistics: (a) frequentist approach, and (b) the Bayesian approach.

Frequentist: if you give an event many opportunities to occur, the probability of occurrence is the

$$\frac{\#\text{ of occurrences}}{\#\text{ of opportunities}} \tag{62}$$

This approach works well when you can repeat an experiment many, many times.

Bayesian: this approach is named after the frequent use of Bayes theorem, which takes into account a priori information that may not be useable by a frequentist.

Often, these two approaches can give different answers:

> For example, in the Monty Hall problem, we wanted to know the probability that the car was behind a certain door, say door C. A frequentist approach would be to say, if I run this experiment thousands of times, how often will the car be behind door C? The answer is 1/3 of the time.
>
> In our Bayesian formulation, we used additional a priori information, namely, that Monty Hall showed us that it wasn't behind a certain door. A frequentist approach does not use this information, while a Bayesian approach does.

> For the cab accident example, a frequentist would say the probability of the cab in the accident being Blue is 80%, the probability the witness was correct. The Bayesian approach, however, also took into account the background rate of the number of cars in the city, information that the frequentist approach couldn't use.

Bayes Theorem is actually used quite often in atmospheric science in the form of the Kalman Filter, for example, for forecasts. At its base level, the Kalman Filter is an iterative application of Bayes Theorem which uses uncertainties in measurements to weigh their use when updating the next time step.

1.3 Statistical Significance Testing

1.3.1 Normal distribution

The probability of an event occurring between a and b is:

$$\boldsymbol{Pr}(a \leqslant x \leqslant b) = \int_a^b f(x)dx, \tag{63}$$

where $f(x)$ is the probability density function (PDF). Note that for continuous $f(x)$:

$$\boldsymbol{Pr}(x = c) = 0 \tag{64}$$

$f(x)$ must have the following characteristics to be a probability density function:

- $f(x) > 0$ for all x
- $\int_{-\infty}^{\infty} f(x) = 1$

The cumulative density function (CDF) at value b is denoted $F(b)$ and is the probability that x assumes a value less than b:

$$F(b) = \int_{-\infty}^{b} f(x)dx \tag{65}$$

$$\text{hence,} \tag{66}$$

$$\boldsymbol{Pr}(a \leqslant x \leqslant b) = F(b) - F(a). \tag{67}$$

Example: PYTHON NOTEBOOK WHAT_IS_PDF.IPYNB

The probability density function for a variable x that is normally distributed about its mean is given by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)} \tag{68}$$

The associated cumulative distribution function is given as:

$$F(b) = \int_{-\infty}^{b} \frac{1}{\sigma\sqrt{2\pi}} e^{\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)} dx \tag{69}$$

The probability that x will fall between two values a and b is thus:

$$F(b) - F(a) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)} dx \tag{70}$$

Often, one standardizes the normal distribution by defining a variable z:

$$z = \frac{x - \mu}{\sigma} \tag{71}$$

Note that in this case, the mean of $z = 0$ and the standard deviation of $z = 1$. This is very useful for discussing properties of the normal distribution with others who may have variables with different means and standard deviations. For example, recall in the discussion of skewness and kurtosis, the standard normal has a skewness of 0 and a kurtosis of 3.

If one standardizes to z, then

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right) \tag{72}$$

(note that the σ no longer appears in the denominator due to the transformation of F(z) using $z = (x - \mu)/\sigma$ and needing to replace dx with $\sigma \cdot dz$)

Example: Draw normal distribution - discuss histograms

$$\boldsymbol{Pr}(-1 \leqslant z \leqslant 1) = 68.27\% \tag{73}$$
$$\boldsymbol{Pr}(-2 \leqslant z \leqslant 2) = 95.45\% \tag{74}$$
$$\boldsymbol{Pr}(-3 \leqslant z \leqslant 3) = 99.73\% \tag{75}$$
$$\tag{76}$$

Recall that z is a standardized normal variable, and so, $z = 2$ is $z = 2$ standard deviation.

Thus, there is a 4.55% probability that z will fall outside of 2 standard deviations of its mean (two-tailed probability) and a 2.275% change it will exceed +2 standard deviations (one-tailed probability).

Example: Draw meaning of one-tailed and two-tailed probabilities.

Example: show uniform distribution example, integrate a rectangle

1.3.2 Statistical Significance Testing: the t-statistic and z-statistic

Many geophysical variables are normally distributed - so often we can use the normal distribution to test if our sample is different from the population mean. However, always check that our data is normal! Rain rate, for example, is not!

Example: Introduce z-table - hand-out table to class

1.3.2.1 Using the z-score

Assume the you have 50 years of monthly ENSO indices (Z), and that the index is well approximated by a standard normal:

1. Using the table, what is the probability that a randomly chosen month has an ENSO index of exactly +2?
   0%

2. Using the table, what is the probability that a randomly chosen month has an ENSO index of +2 or greater?
   1-.9772 = 2.28%

3. What is the probability the ENSO index is 1.23 or smaller?
   0.8907 = 89.07%

4. What is the probability the ENSO index is 1.23 or greater?
   1-0.8907 = 10.93%

5. What ENSO index ϵ gives $\boldsymbol{Pr}(\mathsf{Z} \geqslant \epsilon) = 0.1$?
   1.285

6. What ENSO index ϵ gives $\boldsymbol{Pr}(\mathsf{Z} \leqslant \epsilon) = 0.9$?
   1.285

7. What ENSO index ϵ gives $\boldsymbol{Pr}(\mathsf{Z} \leqslant \epsilon) = 0.1$?
   -1.285, use symmetry

8. What ENSO index ϵ gives $\boldsymbol{Pr}(|\mathsf{Z}| \geqslant \epsilon) = 0.1$?
   1.645 and -1.645

Some useful values,

$$\boldsymbol{Pr}(-1.96 \leqslant z \leqslant 1.96) = 95\% \quad (77)$$
$$\boldsymbol{Pr}(-2.58 \leqslant z \leqslant 2.58) = 99\% \quad (78)$$
$$(79)$$

Example: PYTHON NOTEBOOK Z-T-CODING.IPYNB

1.3.3 Estimating the statistical significance of the sample mean with the z-statistic

Thus far, we have covered how to determine the probability of getting a value x_i within a range $[a\ b]$, however, in geophysics, we tend to me more interested in the difference between a *sample mean* and an underlying population.

We rarely ask, "what is the probability the ENSO index was its value or greater this month?"

Instead, we might ask, "Was the ENSO index in 2013 consistent with the climatological behaviour?"

Let $X = [x_1, x_2, x_3, x_4, ..., x_N]$ be a random sample of size N drawn from a normal distribution with population mean μ and standard deviation σ.

For $N > 30$, $\overline{X}$ is normally distributed with,

$$\mu_{\overline{X}} = \mu = \text{mean} \quad (80)$$
$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{N}} = \text{standard deviation} \quad (81)$$

In other words, the sample mean is an estimate of the population mean, and the error of your estimate (the variance of the distribution of $\overline{X}$) decreases as N increases (as it should).

$\sigma_{\overline{X}}$ is known as the standard error of the mean.

So, how do we test whether the mean of the sample is different from the mean of the population? We can use the z-statistic!

Recall,

$$z = \frac{x - \mu}{\sigma} \quad (82)$$

Now, we want to test $\overline{X}$, not x. So, replace μ with the population mean of $\overline{X}$ and σ with the population standard deviation of $\overline{X}$.

Plugging things in leads to

$$z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{N}}} \quad (83)$$

The z statistic is now the number of standard errors that the sample mean deviates from the population mean.

We can manipulate the equation for the z-statistic to obtain an equation for the difference of two means (as opposed to the difference between a sample mean and the population), where σ_1 can be different from σ_2:

$$z = \frac{(\overline{x_1} - \overline{x_2}) - \Delta_{1,2}}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}} \tag{84}$$

$\Delta_{1,2}$ is the hypothesized difference between the two means, which is typically 0 in practice.

1.3.3.1 Statistical significance of the mean

(a) What is the probability that December 2013 had an ENSO index of 0.50 or greater?

(b) What is the probability that the average 2011-2013 monthly ENSO index was 0.50 or greater assuming that ENSO dynamics have not changed?

(a) Assume the ENSO index is standard normal. $\mu = 0, \sigma = 1$. Look at table for $z = 0.50$. **1-.69 = 31.0%**. So, not very rare.

(b) Here, we are testing for the sample mean.

$$\overline{X} = 0.5 \tag{85}$$

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{N}} = \frac{1}{\sqrt{36}} \tag{86}$$

$$z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{0.50 - 0}{.1667} = 3.0 \tag{87}$$

The $\boldsymbol{Pr}(z \geqslant 3.0) \approx 1 - 0.9987 = 0.1\%$.

Such a low probability implies that we had either a very rare event, or, that the dynamics of ENSO changed in 2011-2013 compared to climatological ENSO variability.

1.3.4 Confidence Interval

1.3.4.1 Example: Calculating confidence limits

Say we have temperatures for 30 winters with a mean of 10° C and a standard deviation of 5° C. What is the 95% confidence interval on the true population mean? You may assume that the temperatures are normally distributed.

$$z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{N}}} \tag{88}$$

$$\mu = \overline{x} - z\frac{\sigma}{\sqrt{N}} \tag{89}$$

Since we are interested in the spread about the mean, and since the normal distribution is symmetric about the mean, this can be rewritten as,

$$\mu = \overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{N}} \tag{90}$$

Since 95% of the population lies within $z_{\alpha/2} = \pm 1.96$, it follows that the confidence limit for μ is

$$\mu = 10.0 \pm 1.96\frac{5.0}{\sqrt{30}} = 10.0^{\circ} \pm 1.7^{\circ} \tag{91}$$

$$8.3^{\circ} \leqslant \mu \leqslant 11.7^{\circ} \tag{92}$$

The 95% confidence interval is defined as the the interval that contains the true parameter 95% of the time.

Example: Pass-out handout called Confidence_intervals.pdf

1.3.5 Central Limit Theorem

The arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed, with standard error of $\sigma/\sqrt{N}$, where N is the length of each sample.

This theorem is the basis of most of what we do in statistics, and most of what we do when assessing the significance of geophysical signals.

Example: PYTHON NOTEBOOK CENTRAL_LIMIT_THEOREM.IPYNB

What the central limit theorem says is that if you have a sample that is large enough, you can use the normally distributed z-statistic to estimate probabilities of getting that mean - no matter the distribution of the underlying data.

1.3.5.1 Example: Rain rate

Rain rates measured at minute-intervals are lognormal. Thus, you *cannot* use the z-statistic to determine, say, the probability of getting a rain rate at any given time of 2 mm/sec. or higher.

However, if you want to know the probability of having a monthly average rain-rate of 2 mm/sec. or higher, you can use the z-statistic!

Note, because of the Central Limit Theorem, this is often all people remember from their statistics course - however, be careful, it does not always apply.

1.3.6 The t-statistic: when the sample size is small

Consider the z-statistic for the sample mean:

$$z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{N}}} \tag{93}$$

To apply this formula, we need to know the σ of the underlying distribution. But often, we don't know what the underlying distribution is. As we showed in the previous example, if we have a large enough sample, the sample standard deviation s is a good approximation of the true population σ.

However, if $N < 30$, this is not the case, and the z-statistic does not apply!

In this case, one must use the Student's t-statistic, introduced by William Sealy Gosset in 1908 to monitor Guinness quality at the brewery in Dublin, Ireland. At the time, he was not allowed to publish anything that he developed while at Guinness, so he used the pseudonym "Student".

The t-statistic is analogous to the z-score, except it also requires:

$$\nu = \text{degrees of freedom} \tag{94}$$
$$\sqrt{N} \rightarrow \sqrt{N-1} \tag{95}$$

The t-statistic is defined as

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{N-1}}}, \tag{96}$$

where s is the sample standard deviation.

Recall from the Central Limit Theorem example that when $N \gtrsim 30$, the true standard deviation of the sample means is well approximated by $s/\sqrt{N}$, and so we can write σ.

However, in this case, we keep the nomenclature s since it is unknown what the true σ is.

Note the similarity with the z-statistic:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{N-1}}} = \frac{\bar{x} - \mu}{\frac{\hat{s}}{\sqrt{N}}} \tag{97}$$
$$\hat{s} = s\sqrt{\frac{N}{N-1}} \tag{98}$$

$\hat{s}$ is the estimate of the standard deviation based on a limited sample.

If we draw a sample of size N from a normally distributed population of mean μ, we find that t is distributed with the following population density:

$$f(t) = \frac{f_0(\nu)}{\left(1 + \frac{t^2}{\nu}\right)^{\left(\frac{\nu+1}{2}\right)}}, \tag{99}$$

$f_0(\nu)$ is chosen as a normalization factor to make $\int_{-\infty}^{\infty} f(t)dt = 1$.

> ν is the number of degrees of freedom $= N-1$. More on independent samples later.

Example: PYTHON NOTEBOOK Z-T-CODING.IPYNB

Comparisons between the t-statistic and z-statistic:

- most often, the t-distribution is the PDF of the sample means *drawn from a normal distribution*, but with small N
- unlike the z-stat, the t-stat depends on N through the degrees of freedom ν
- smaller values of N lead to longer tails for the t-stat

- as N increases, the t-stat approaches the z-stat
- the key difference is that the t-stat uses an estimate of the standard error based on the *sample* standard deviation s, instead of the true standard error σ

Confidence intervals for the t-statistic work similarly to the z-statistic,

$$\mu = \overline{x} \pm t_c \frac{s}{\sqrt{N-1}} \tag{100}$$

t_c is the critical value for t. It depends on the sample size and the significance level desired. You can see values of this statistic in the t-table. Note that it is setup differently than the z-table since it requires a column for ν.

The difference of means for the t-stat. is very similar to that for the z-stat, but with slight modifications. Assume two samples N_1 and N_2 are drawn from normal distributions who's standard deviations are equal ($\sigma_1 = \sigma_2$).

$$t = \frac{(\overline{x_1} - \overline{x_2}) - \Delta_{1,2}}{\hat{\sigma}\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \tag{101}$$

$$\hat{\sigma} = \sqrt{\frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}} \tag{102}$$

where $\nu = N_1 + N_2 - 2$ and $\Delta_{1,2}$ is the hypothesized difference.

The pooled variance $\hat{\sigma^2}$ is a weighted average of the sample variances.

A similar statistics applies for the z-score, but we use the true population variances instead:

$$z = \frac{(\overline{x_1} - \overline{x_2}) - \Delta_{1,2}}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}} \tag{103}$$

1.3.6.1 Example: Comparing the z-stat and t-stat

You have 5 years of monthly-mean temperature data derived from the MSU4 satellite. The mean temperature along 60N during January is $\sim -60^{o}$ C and the standard deviation is $\sim 8^{o}$ C. What are the 95% confidence limits on the true population mean?

---

z-statistic

The critical value $z_c = \pm 1.96$ for 95% confidence. Thus, the population mean μ is expected to lie within

$$-60 \pm 1.96\frac{8}{\sqrt{5}} = -60 \pm 7.0 \rightarrow -67.0 \leqslant \mu \leqslant -53.0 \tag{104}$$

t-statistic

The critical value $t_c = \pm 2.78$ for $\nu = 5 - 1 = 4$. (Note: we want to use 0.025 since a two tailed test). Thus, the population mean μ is expected to lie within

$$-60 \pm 2.78\frac{8}{\sqrt{4}} = -60 \pm 11.1 \rightarrow -71.1 \leqslant \mu \leqslant -48.9 \tag{105}$$

Thus, using the t-statistic gives a wider confidence range than the z-statistic, reflecting the additional uncertainty associated with a small N. If we had erroneously used the z-stat instead of the t-stat, we would underestimate the 95% confidence bounds by 35%.

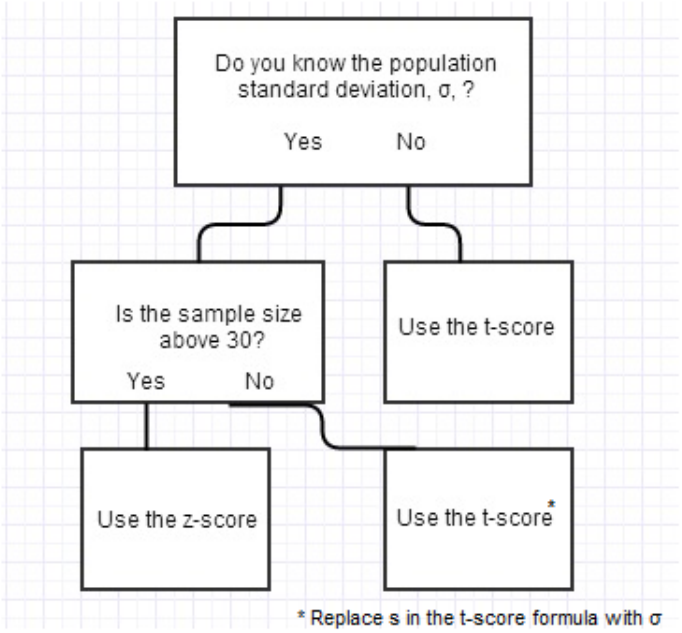

Figure 1: When to use the z-score vs the t-score.

1.3.7 When to use (and not use) the t-test

If you are choosing between the z-test and t-test, use the t-statistic! It converges to the z-test for large N.

When applying the t-test to compare means of two samples, you are implicitly making a very strong assumption: *that the underlying distributions of both samples are normal.*

ALWAYS CHECK YOUR NORMALITY ASSUMPTION: plot your data, look at the skewness, kurtosis, etc. There are fancy tests to check for normality, e.g. the Kolmogorov-Smirnov test.

Now, we previously discussed the Central Limit Theorem, which tells us that for a "large enough" sample size, the distribution of sample means is normal. HOWEVER, note that the t-test applies for small N but for underlying *normal* distributions, whereas the Central Limit Theorem only applies for large N.

Thus, you cannot blindly apply the t-test to test differences in sample means if the underlying distributions are not already normal (although there are limited exceptions). This is a common mistake made in our field.

Example: PYTHON NOTEBOOK CENTRAL_LIMIT_THEOREM.IPYNB

1.3.8 A note on the independence on N

In all of the cases thus far, it has been assumed that the N samples are *independent* samples. Often, N observations of a geophysical variable are not independent, for example, they exhibit either spatial or temporal correlations.

For example, the geopotential height is highly auto-correlated so that each day's value is not independent from the previous or following days. You cannot improve the your ability to know a 5-day wave by sampling every 3 hours instead of every 6. We will discuss this more later-on in the course.

1.4 Hypothesis Testing

1.4.1 Terminology and symbology

Example: DRAW PICTURE OF α, P-VALUE, t_c, ETC.

- significance/confidence level: α, typically 5% (0.05), often reported as $1 - \alpha$
- critical value: t_c or z_c, the value that must be exceeded to reject the null hypothesis, one-sided t_α, two-sided $t_{\alpha/2}$
- p-value: probability of observing an effect given that the null hypothesis is true (probability of your t-score or z-score)

1.4.2 Setting-up the problem

In using statistical significance tests, there are 5 basic steps that should be followed, *in order*.

1. State the significance level (α)
2. State the null hypothesis H_0 and the alternative H_1
3. State the statistic to be used, and the assumptions required to use it
4. State the critical region
5. Evaluate the statistic and state the conclusion

Proper construction of the null hypothesis and its alternative is critical to the meaning of statistical significance testing.

Usually the null hypothesis is a rigorous statement of the conventional wisdom or a zero information conclusion, and its alternative is an interesting conclusion that follows directly and uniquely from the rejection of the null hypothesis.

Some examples:

H_0: The means of two samples are equal.
H_1: The means of two samples are not equal.

H_0: The anomaly is zero.
H_1: The anomaly is not zero.

H_0: The correlation coefficient is zero.
H_1: The correlation coefficient is not zero.

Hypothesis testing is much weaker than Bayesian statistics (discussed at the beginning of this course). All you are really doing is *stating whether the data is consistent with the null hypothesis or not.*

You are not saying that the null hypothesis is true, or that the alternative is true, or that either is false.

1.4.2.1 Example of Hypothesis Testing

Sam went skiing 10 times at Vail this winter. The average temperature on these 10 days was 35F, and the standard deviation of 10 daily temperatures is 5F. Sam knows that the climatological mean winter temperature for Vail is 32F. Is this a sign of climate change?

Let's suppose by a "sign of climate change", we mean "Is the 10-day average experience by Sam consistent with the null hypothesis that Vail temperatures have not warmed?" Let's walk through the 5-steps to answer this question.

Let μ be the true mean of the population from which the 10 days Sam went skiing were sampled.

1. let's use 95% confidence ($\alpha = 0.05$)

2. H_0: $\mu = 32^o$
   H_1: $\mu \neq 32^o$

3. For this problem, we will assume that the temperatures at Vail are normally distributed. Since we have $N = 10$ samples to estimate the true standard deviation (σ) from the sample standard deviation (s), we must use the t-test.

4. We will use a two-sided t-test (before he went skiing, Sam didn't know what to expect from the temperatures). Thus, to reject the null hypothesis we must have $t > t_{0.025} = 2.262$ (for $\nu = 9$)

5. $t = \frac{35-32}{\frac{5}{\sqrt{10-1}}} = 1.80$

   $t = 1.80 < t_{0.025}$, so we cannot reject the null hypothesis that the underlying population is different for the days Sam went skiing.

   Note, we should not say:

   > "There was nothing different about the temperatures on the day Sam went skiing."
   > *...nor...*
   > "There has not been a change in temperatures at Vail over the past few decades."

   We can only say

   > "The data is consistent with the null hypothesis that the mean of the underlying population on the days Sam went skiing was the same as the climatology."
   > *...or...*
   > "The higher temperatures present when Sam went skiing are consistent with natural variability."

1.4.2.2 Comparison of Means & Hypothesis Testing

In 2000, you and your science team went into Rocky Mountain National Park to measure the Nitrogen deposition there. You took 17 samples and found they had an sample mean of 30 $\mathrm{mg} \cdot \mathrm{N/m^2}$, and a sample standard deviation of 10 $\mathrm{mg} \cdot \mathrm{N/m^2}$.

In 2014, you repeat your measurements, but this time, due to budget cuts, you are only able to take 7 samples, resulting in a sample mean of 45 $\mathrm{mg} \cdot \mathrm{N/m^2}$ and a sample standard deviation of 20 $\mathrm{mg} \cdot \mathrm{N/m^2}$.

Has Nitrogen deposition significantly increased between 2000 and 2014?

This problem is different than one we have done before, since we are not comparing our sample to a long-term average (or climatology), but we are comparing two sample means to each other. Thus, we need to use "a comparison of sample means" test. The good news is that it is similar to what we have been doing thus far.

Assume two samples N_1 and N_2 are drawn from normal distributions who's true standard deviations are equal ($\sigma_1 = \sigma_2$; if they are not assumed equal, you want to use a different test called *Welch's t-test*). Then, the t-statistic for sample means is

$$t = \frac{(\overline{x_1} - \overline{x_2}) - \Delta_{1,2}}{\hat{\sigma}\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \tag{106}$$

$$\hat{\sigma} = \sqrt{\frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}} \tag{107}$$

where $\nu = N_1 + N_2 - 2$ and $\Delta_{1,2}$ is the hypothesized difference. The "pooled variance" $\hat{\sigma}^2$ is a weighted average of the sample variances.

A similar statistic applies for the z-score, but we use the true sample variances instead and they do *not* have to be equal to each other:

$$z = \frac{(\overline{x_1} - \overline{x_2}) - \Delta_{1,2}}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}} \tag{108}$$

Now, let's answer the question using hypothesis testing:

1. 95% confidence ($\alpha = 0.05$)
2. H_0: $\mu_{2000} = \mu_{2014}$
   H_1: $\mu_{2000} \neq \mu_{2014}$
3. We will use the t-statistic for the comparison of means and we will assume that (1) the variance of Nitrogen deposition has not changed and (2) the underlying distribution from which your Nitrogen samples were drawn follows a normal distribution (we can check this). We will use a two-sided test since we didn't know a priori if the difference would be positive or negative.
4. We will reject the null hypothesis if $t \geqslant t_c = t_{0.025} = 2.0739$ (for $\nu = 17 + 7 - 2 = 22$).
5. Evaluate

$$\hat{\sigma} = \sqrt{\frac{17 \cdot 10^2 + 7 \cdot 20^2}{17 + 7 - 2}} = 14.3 \tag{109}$$

$$t = \frac{45 - 30 - 0}{14.3\sqrt{\frac{1}{17} + \frac{1}{7}}} = 2.36 \tag{110}$$

   Since $t = 2.36 > t_c$, we can reject the null hypothesis that the sample means in the two years are equal.

1.4.3 Type I and Type II errors

	H_0 is true	H_0 is false
Fail to Reject H_0	No Error	Type II Error (false negative)
Reject H_0	Type I Error (false positive)	No Error

The way typical hypothesis tests (frequentist approach) are setup, a 95% confidence level means you have a 5% chance of making a *Type I Error*, that is, you reject the null hypothesis (think you found something interesting) when you should not have. It is much more difficult to asses the Type II Error - the probability you "play it safe and fail to reject H_0 when something interesting was there". For typical hypothesis testing, the probability or a Type II error can be very large.

In engineering, you often care about the differences between Type I and Type II errors, and you design your statistics to reflect your judgements. For example, if H_0 is that the bridge will hold-up if 10 semi-trucks cross at the same time, and H_1 is that the bridge will not hold-up, you might be happier with a Type I Error, which requires that you redesign the bridge, rather than a Type II Error, where you think the bridge will be fine, and it won't be.

1.4.4 A priori vs. A posteriori

A priori: you have a reason to expect a relationship ahead of time.
A posteriori: you don't. This distinction is critical for hypothesis testing.

One-tailed vs. Two-tailed tests:
If you have an "a priori" expectation of the sign of the result, you can use a 1 tailed test. Otherwise, you should use a two tailed test.

So far, the specific significance levels we have discussed rely on a priori statistics. Here are some examples,

Example: PYTHON NOTEBOOK MAPS_OF_RANDOM_RELATIONSHIPS.IPYNB

1.4.4.1 Gone fishin' I: a priori vs a posteriori statistics

The jelly bean example on the syllabus (xkcd) shows students testing whether 20 different colors of jelly-beans may cause acne, with a significance level of 95% confidence. Let H_0: jelly beans do not cause acne. What is the probability that 1 jelly-bean color would incorrectly reject H_0 and find a "significant result" by chance alone?

$$\boldsymbol{Pr}(\text{you do not reject } \mathsf{H}_0 \text{ when } \mathsf{H}_0 \text{ is true for 1 test}) = 0.95 \quad (111)$$

$$\boldsymbol{Pr}(\text{you do not reject } \mathsf{H}_0 \text{ when } \mathsf{H}_0 \text{ is true for 20 tests}) = 0.95^{20} = 36\% \quad (112)$$

So, our 95% test is really a 36% - we have a 64% chance of finding at least one relationship by chance!

1.4.4.2 Gone fishin' II: a priori vs a posteriori statistics

You think that Arctic warming has caused blocking to increase in frequency between the 1980's and today. You look over the 4 seasons, two latitude bins (north of 45° N and south of 45°N), and 6 different longitude bins. You test for changes in mean blocking frequency at the 95% confidence level. How many "significant changes" in blocking should you expect by chance alone? How might you apply a posteriori statistics?

You have no a priori knowledge which season or longitude bin should exhibit changes in blocking due to sea ice - thus, you are giving the test $4 \times 6 \times 2 = 48$ chances to succeed. The number of changes you expect to show significant trends by chance alone is $0.05 \times 48 = 2.4$, so, you should find 2.4 "significant changes" by chance alone.

Let H_0 : the mean in blocking frequency has not changed.

$$\boldsymbol{Pr}(\text{correctly not reject } \mathsf{H}_0 \text{ when it is true for one test}) = 0.95 \quad (113)$$

$$\boldsymbol{Pr}(\text{correctly not reject } \mathsf{H}_0 \text{ when it is true for all 48 tests}) = 0.95^{48} \approx 9\% \quad (114)$$

Thus, your 95% confidence level is really a 9% confidence level!

By trial and error, we can calculate the significance level β for which $\beta^{48} \approx 0.95$ (our a posteriori statistic). In this case, $\beta \approx 0.999$. Thus, if we require the blocking changes for each chance to pass at the 99.9% confidence level, then the probability of correctly not rejecting the null hypothesis for all chances will be 95%.

1.4.4.3 Bayesian vs Frequentist approach

Bayesian probability:

- the evidence about the true state of the world is expressed in terms of degrees of belief
- any probability is a conditional probability given what one knows (varies from person to person)
- probability is seen as the "plausibility of an outcome"
- "updates" a prior hypothesis when new information comes to light
- often answers the question of "how likely is it that a hypothesis is true given my data?", that is,

$$\boldsymbol{Pr}(\text{hypothesis} \mid \text{data})$$

Frequentist probability/hypothesis testing:

- any given experiment can be considered as one of an infinite sequence of possible repetitions
- result of a frequentist approach is often a "true or false" conclusion from a significance test
- probability is seen as the "potential frequency of an outcome"
- often answers the question "how likely is my data if the null hypothesis is true?", that is,

$$\boldsymbol{Pr}(\text{data} \mid \text{null hypothesis})$$

1.4.4.4 Bayesian vs Frequentist approach cont...

Example: SEE BAYESIAN_VS_FREQUENTIST_COSMICRAYS.IPYNB

You've heard the hypothesis that variations in cosmic rays are driving the increase in global-mean temperatures we've recently witnessed, but you are very skeptical. You've been to a lot of seminars by experts in the field suggesting there is little evidence for this hypothesis. You calculate the correlation between incoming cosmic ray intensity and global-mean temperature and the resulting correlation has a p-value of 0.05. Should you now accept that cosmic rays are the answer and forget about CO_2?!

Frequentist

- H_0 : there is no real relationship between cosmic rays and global temperature increases
- H_a : there is a real relationship between cosmic rays and global temperature increases
- p-value $= 0.05$
- conclusion: if your confidence-level is $\alpha = 0.05$, then you conclude that you can "reject the null-hypothesis" that there is no relationship between cosmic rays and the global temperature increase. Note that this approach doesn't care about H_a, nor does it care about all the evidence against cosmic rays.

Bayesian

- R : cosmic rays are driving the increase in global mean temperature
- $\tilde{\mathsf{R}}$: cosmic rays are not
- E : the evidence obtained from the data, e.g. the correlation
- Want: $\boldsymbol{Pr}(\mathsf{R}|\mathsf{E})$
- Know: $\boldsymbol{Pr}\left(\mathsf{E}|\tilde{\mathsf{R}}\right) = 0.05$
- Bayes' Theorem:

$$\boldsymbol{Pr}(\mathsf{R}|\mathsf{E}) = \frac{\boldsymbol{Pr}(\mathsf{E}|\mathsf{R})\,\boldsymbol{Pr}(\mathsf{R})}{\boldsymbol{Pr}(\mathsf{E}|\mathsf{R})\,\boldsymbol{Pr}(\mathsf{R}) + \boldsymbol{Pr}\left(\mathsf{E}|\tilde{\mathsf{R}}\right)\boldsymbol{Pr}\left(\tilde{\mathsf{R}}\right)} \tag{115}$$

- Need to Know: $\boldsymbol{Pr}(\mathsf{E}|\mathsf{R})$, $\boldsymbol{Pr}(\mathsf{R})$ [note that $\boldsymbol{Pr}\left(\tilde{\mathsf{R}}\right) = 1 - \boldsymbol{Pr}(\mathsf{R})$]
- Expert Opinion: $\boldsymbol{Pr}(\mathsf{R})$ is your prior probability, what you thought the likelihood that cosmic rays were driving the temperature changes *before* you analyzed the data. Since you were pretty skeptical due to expert opinion, let's say $\boldsymbol{Pr}(\mathsf{R}) = 0.01$ so $\boldsymbol{Pr}\left(\tilde{\mathsf{R}}\right) = 1 - .01 = .99$.
- Expert Opinion: We also need to know $\boldsymbol{Pr}\,(\mathsf{E}|\mathsf{R})$, that is, the probability of getting the correlation you did if cosmic rays *are* causing the temperature changes. This one is tricky. For now, let's say that $\boldsymbol{Pr}(\mathsf{E}|\mathsf{R}) = 0.95$, that is, you believe if there was an influence of cosmic rays, you would very likely calculate a correlation as big (or bigger) than what you actually did.
- Bayes' Theorem:

$$\boldsymbol{Pr}(\mathsf{R}|\mathsf{E}) = \frac{\boldsymbol{Pr}(\mathsf{E}|\mathsf{R})\,\boldsymbol{Pr}(\mathsf{R})}{\boldsymbol{Pr}(\mathsf{E}|\mathsf{R})\,\boldsymbol{Pr}(\mathsf{R}) + \boldsymbol{Pr}\left(\mathsf{E}|\tilde{\mathsf{R}}\right)\boldsymbol{Pr}\left(\tilde{\mathsf{R}}\right)} = \frac{0.95 \cdot 0.01}{0.95 \cdot 0.01 + 0.05 \cdot 0.99} = 0.16 \tag{116}$$

- Conclusion: you give it a 16% likelihood that cosmic rays are driving the global temperature increase. Put another way, the probability of your null hypothesis being true started off at 99% and now it is still 84% (not 5%)! The likelihood is lower than your p-value because you've allowed your prior knowledge to come into the calculation - you thought that the hypothesis was unbelievable (a "rare event") to begin with. Note that as $\boldsymbol{Pr}(\mathsf{E}|\mathsf{R})$ decreases, the $\boldsymbol{Pr}(\mathsf{R}|\mathsf{E})$ decreases as well. For example, if $\boldsymbol{Pr}(\mathsf{E}|\mathsf{R}) = 0.5$ (instead of 0.95) you get a likelihood of 9%. If you had no prior knowledge, you would let $\boldsymbol{Pr}(\mathsf{R}) = \boldsymbol{Pr}\left(\tilde{\mathsf{R}}\right) = 0.5$, and you get 95%!

1.4.4.5 Another Bayesian example...

Example: PYTHON NOTEBOOK BAYESIAN_SST_EXAMPLE.IPYNB

You are interested in measuring the sea-surface temperature off of the Washington coast to within 0.1 units. You know from previous years that the distribution of anomalous sea-surface temperature (denoted as random variable X) follows a uniform distribution with bounds $(-2, 2)$. The problem is that the instrument you use to measure the temperature is known to introduce an additive error. While the exact error during each measurement is unknown, you know from calibration that the error follows a normal distribution with $\mu = 0$ and $\sigma = 0.75$, denoted at W. That is, you measure a random variable Y that is the sum of the two random variables X and W: $Y = X + W$.

You go out and take a measurement and the instrument says that the anomalous sea-surface temperature is $y = 2.2$. (a) What is the probability that $X = x \pm \epsilon$ given that you measured y, or in math, $\boldsymbol{Pr}(X = x \pm \epsilon | Y = 2.2 \pm \epsilon)$ for some range determined by ϵ? (b) For each possible value of y, what is your best estimate of x?

Bayesian

(a) Fig. **??** shows the distributions of X, W and Y. Using Bayes' Theorem, you want to compute the following:

$$\boldsymbol{Pr}(X = x \pm \epsilon | Y = y = 2.2 \pm \epsilon) = \frac{\boldsymbol{Pr}(Y = y \pm \epsilon | X = x \pm \epsilon)\, \boldsymbol{Pr}(X = x \pm \epsilon)}{\boldsymbol{Pr}(Y = y \pm \epsilon)} \tag{117}$$

where ϵ is determined by you. Here, we will let $\epsilon = 0.05$.

You know that X follows a uniform distribution, and the distribution of Y given $X = x$ follows the normal distribution of W but centered on $\mu = x$ instead of zero. Determining the distribution of Y is a bit trickier. What you want to know is the distribution of the sum of a normal distribution and a uniform distribution - however, this is not simple to write down analytically. One can determine this distribution using convolution (to be discussed later in the course), model it empirically, or use a built-in software package (I use python's PaCAL). The distribution of Y is shown in Fig. **??**. Plugging values into Bayes' Formula leads to Fig. **??**.

(b) To answer this, you must decide what you mean by "best estimate". Here, you will define the best estimate based on the posteriori mean. That is, the best estimate of X given $Y = y$ is defined by the *minimum mean squared error estimate*, which is just the expected value of X given $Y = y$. In the case of $Y = y = 2.2$, the expectation of X ($x = 1.46$) is shown by the gray line in Fig. **??**.

Performing this calculation for all possible values of y leads to Fig. **??**, which shows the expectation of sea-surface temperature given that you measure a particular value of y. Note how the curve slowly approaches -2 and 2 since X is bounded by these two values.

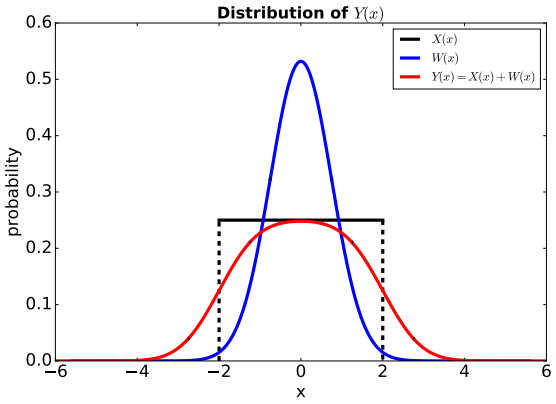

Figure 2: Probability density functions of X, W *and* $\mathsf{Y} = \mathsf{X} + W$.

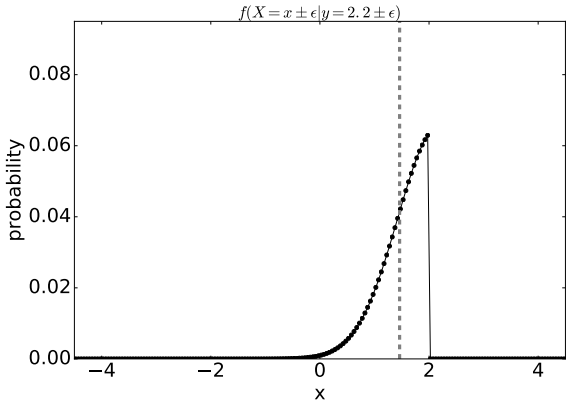

Figure 3: The probability that $\mathsf{X} = \mathsf{x} \pm \epsilon$ *given that* $\mathsf{Y} = \mathsf{y} = 2.2 \pm \epsilon$. *The dashed gray line denotes the conditional expectation of* X *based on* Y.

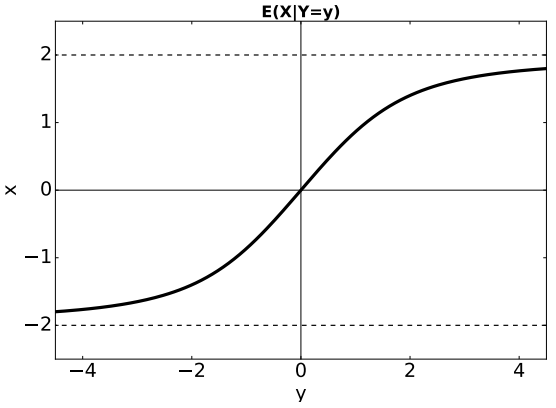

Figure 4: The conditional expectation of X *given* $\mathsf{Y} = \mathsf{y}$.

1.5 Monte Carlo and Resampling techniques

Monte Carlo and resampling techniques haven't historically been discussed or used as often because of the computational constraints. These days, computer time is cheap and everyone has one, so, these techniques can be found more regularly in the peer reviewed literature.

Resampling techniques: best used when you have a long climatology or control run

Monte Carlo: useful when you think you know the underlying distribution or behaviour, but you don't have a long control or climatology - thus, you create many synthetic "realities" to determine the probabilities of certain events

1.5.1 Why use resampling and Monte Carlo techniques?

1. when spatial and temporal correlations are hard to estimate
2. it is unclear which statistical assumptions are appropriate (e.g. your underlying population is not Gaussian, and the central limit theorem does not apply) - this is when resampling techniques are useful
3. the statistic of interest does not have a straight-forward theory for estimation, but you know the underlying population's distribution - this is when Monte Carlo simulations are useful

1.5.2 Resampling: bootstrap

Bootstrapping involves constructing a number of resamples of the original dataset (and of equal size to the observed sample of interest) by random sampling with replacement from the original dataset. In this way, you never need to assume anything about the underlying distribution of the data since it is already built-in to the original dataset. In essence, you ask, by random chance, what is the probability that a particular event (or sample statistic) occurred?

This method is also useful when you are determining statistics other than the mean (e.g. extrema, median, skewness) when we don't have simple statistics for these variables.

A reference to the bootstrap method is: Efron, B., "Bootstrap Methods: Another Look at the Jackknife", The Annals of Statistics, Vol. 7, No. 1, pp. 1-26.

The technique is called "bootstrapping" because it almost appears you get something out without putting something in. This is a phrase often used in computer programming to refer to a small amount of simple software that can load more complex software that loads more complex software (etc., etc.,) almost as if the program is "pulling itself up by its bootstraps."

1.5.3 Resampling: jacknife

The jackknife method predates the bootstrap method. It is a way of getting errors bars (or measuring the variance) of your sample. You systematically remove one value from your sample, and calculate the statistic, then put the value back into the sample and remove the next value, calculate the statistics...and on and on. Then, following a simple formula, you can estimate the variance of the parameter by using the mean of the jackknife estimates of your sample parameter of interest.

You will find the jackknife approach often used to estimate the slope of a line through multiple points, or the y-intercept, etc.

Example: RUN NOTEBOOK JACKKNIFE_EXAMPLE.IPYNB

1.5.4 Monte Carlo

Monte Carlo simulations require that you make an assumption about the underlying distribution. This is useful when you don't have a large enough base population to perform the bootstrap approach.

The idea is that you create bunch of synthetic (i.e. fake) data according to your distribution - you create many "realities" based on the same underlying distribution, and then you perform your analysis on these realities to determine the distribution of the statistical parameters of interest.

For example, in Homework 1, you wrote Monte Carlo simulations using the underlying normal distribution.

1.5.4.1 Resampling: bootstrap

You believe that aerosols grow the most when you have high geopotential heights nearby. You composite the 500 geopotential height on the 20 August days when you have aerosol formation and growth over a site in Egbert, Canada, and you find that the average geopotential on these days is 5900 m. The mean at this station is 5886 m, so the heights are higher. Are these results significant? Or is this just random chance?

Run 2,500 experiments, within each experiment, randomly grab 20 days from the historical geopotential height data, and take the mean of the 20 days. After 2,500 iterations, you will have a distribution of the $\mathsf{N} = 20$ sample means under the null hypothesis of random chance. Now, you can look at this distribution and determine the 95% confidence bounds on the $\mathsf{N} = 20$ sample means - if the observed value of 5900 m outside of this range, you have reason to believe it may be more than random chance.

Example: RUN SUBSAMPLING_EXAMPLE.PY

1.5.4.2 Monte Carlo simulation

In January (31 days), the maximum daily temperature was 2.2 standard deviations from the climatological mean temperature. If we assume that the daily temperature is normally distributed, how rare is it to have a maximum of 2.2σ or greater in 31 daily samples?

We do not have a test for the maximum of a distribution - note, *this is not the mean*.

We can't use the bootstrap approach, since we don't have the population to resample from, we only have our 31 points.

However, since our null hypothesis is that the values come from a standard normal, we can create synthetic data to determine the confidence interval on the maximum in a sample of $\mathsf{N} = 31$.

Example: RUN MONTE_CARLO_EXAMPLE.PY

If we had the climatological data (say 50 years of data), we could instead follow a bootstrap approach and resample the full time series many times to get the distribution of the maximum value for samples of length $\mathsf{N} = 31$. In this case, the underlying distribution would not need to be normal (as in the bootstrap example).

1.6 Compositing

Compositing, also sometimes called *superposed epoch analysis* when applied to time series, is one of the simplest analysis techniques imaginable. It is very powerful, but can also be misused.

Compositing: sorting data into categories and comparing the statistics for different categories.

The idea is by averaging the data in a smart way, you can isolate the signal and remove the background "noise" (unwanted/not interesting to you - signals)

Steps to Compositing:

1. Determine categories
   - for diurnal cycle, use time of day
   - impacts of ENSO, warm/cold SSTs
   - impacts of sea ice loss, high and low September sea ice years

   You should have an a priori hypothesis as to why the variable being composited should depend on the category
2. compute the statistics for each category
3. display results
4. validate results
   - calculate relevant statistics (most often z-test using a comparison of means), if N is big enough (i.e. enough data in each category to use the Central Limit Theorem)
   - OR, perform Monte Carlo or sub-sampling techniques
   - subdivide the data and show relationship exists in sub-samples of the data

Advantages over regression (fitting a line): can isolate nonlinear relationships, don't need to make assumptions about the underlying distributions

Disadvantages: does not use all of the data (more susceptible to sampling errors), tends to focus on "extremes" of each category

Example: Slides on compositing: sea ice and ENSO/PSDI

1.6.1 Significance of composites

With composites, you often compare your composited field to its background (climatological) field. If the data if normally distributed, then you can use the t-test when your composite sample is small.

For example, monthly-mean temperatures are relatively normally distributed across the globe. So, in our sea ice example, the composite for low sea ice only had 5 years, so, a t-test with $\nu = 5 - 1 = 4$ will be fine.

When your composites are much larger, then of course, a z-test if just fine.

As mentioned, analyzing only a subset of the data is a good way to make sure your result is robust. However, if $N = 5$, you probably don't have enough data to subdivide the data. If $N = 50$, you could do the analysis in two random $N = 25$ chunks, and confirm you get similar results.

However, what if the underlying distribution isn't normal, or you don't have enough data to subdivide? At this point, you can run into problems. Furthermore, atmospheric data is spatially and temporally correlated, so, you can still get "significant results" according to a t-test by random chance (recall the example from the first lecture).

Example: NOTEBOOK MAPS_OF_RANDOM_RELATIONSHIPS.IPYNB

Example: DISCUSS WILKS (2016; BAMS)

1.7 Other Common Distributions

1.7.1 Chi-square Distribution: tests of variance

Sometimes we want to test if the sample variances are truly different. For this we cannot use t-test or z-test, but we can use the Chi-square distribution.

Before we get to the distribution, let's define a random variable χ^2:

$$\chi^2 = (N - 1)\frac{s^2}{\sigma^2} \tag{118}$$

This quantity can be used to test if the sample variance s^2 is different from the population variance σ^2. Note we are using a ratio, rather than a difference, since a difference ends up being very complicated.

If the underlying distribution from which we draw N values to compute χ^2 is a normally distributed population with standard deviation σ, then the χ^2 values will be distributed as follows:

$$f(\chi^2) = f_0(\nu)(\chi^2)^{(\frac{1}{2}\nu - 1)} \exp^{-\frac{1}{2}\chi^2}, \tag{119}$$

where f_0 is a normalization factor. This is the *Chi-square distribution*. The Chi-square distribution is a member of the Gamma family of continuous probability functions.

The Chi-square distribution can be used to estimate the significance of the ratio $\frac{s^2}{\sigma^2}$.

If you are trying to determine the "true" variance, you can move things around to get the confidence limits for the true variance given your sample variance. That is,:

$$\frac{s^2(N-1)}{\chi^2_{0.975}} \leqslant \sigma^2 \leqslant \frac{s^2(N-1)}{\chi^2_{0.025}} \tag{120}$$

Notes on Chi-square:

- χ^2 is used to assess the confidence limits on σ
- like the t-distribution, the Chi-square distribution is a function of ν
- the Chi-square distribution is not symmetric about its mean: $\chi^2_{0.025} \neq \chi^2_{0.975}$
- for $\nu > 30$, the Chi-square distribution approaches the Normal distribution
- `https://en.wikipedia.org/wiki/Chi-squared_distribution`

Example: NOTEBOOK CHISQUARED_DISTRIBUTION.IPYNB

1.7.2 F-statistic

The F-statistic is used to assess the ratio between two sample standard deviations s_1 and s_2 - whereas the Chi-square is used when you have one sample and you are comparing it to a known population variance σ. In other words, the Chi-square only has a small sample N_1, while the F-statistic has a small sample N_1 and N_2.

If s_1 and s_2 are the variances of independent random samples of size N_1 and N_2, taken from two Normal populations having true variances σ_1^2 and σ_2^2, respectively, then we can define a random variable F as:

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \tag{121}$$

F is a random variable that follows an F-distribution with parameters $\nu_1 = N_1 - 1$ and $\nu_2 = N_2 - 1$. Note that the F-statistic is the ratio of two chi-squared random variables (scaled appropriately due to the N-1).

Note that if both s_1 and s_2 come from Normal populations with the same variances, i.e. $\sigma_1 = \sigma_2$ then the F-statistics simplifies to:

$$F = \frac{s_1^2}{s_2^2} \tag{122}$$

The F-distribution has the following properties:

- the mean is $\frac{\nu_1}{(\nu_2-2)}$
- the variance is $\frac{\nu_2^2(\nu_1+2)}{\nu_1(\nu_2-2)(\nu_2-4)}$

Note that the F-statistic will be used extensively when testing the significance of peaks in frequency spectra (Part III of this course).

Example: NOTEBOOK CHISQUARED_DISTRIBUTION.IPYNB

1.7.3 Binomial

The binomial distribution is often one of the first taught in college statistics. Suppose you have a set of N trials where the outcome of each trial is either a "success" or "failure". These trials are called "Bernoulli trials".

The probability of a success in one trial is $p = \boldsymbol{Pr}(\mathsf{success})$. If X is the total number of successes in N trials, then

$$\boldsymbol{Pr}(X = k) = \binom{N}{k} p^k (1-p)^{n-k} \tag{123}$$

Recall that

$$\binom{N}{k} = \frac{N!}{k!(N-k)!} \tag{124}$$

Note that the right-hand-side looks complicated, but really is the probability of a certain number of successes times the probability of a certain number of failures, with a special factor in-front since the order of successes and failures does not matter.

Example: Binomial examples

1.7.4 Normal Approximation to the Binomial

You can probably imagine the problems getting much more tedious as you have to sum more and more. However, there is a normal approximation to the Binomial distribution in the case of

1. large N
2. $Np \geqslant 10$
3. $N(1-p) \geqslant 10$

In this case, the statistic

$$z = \frac{X - Np}{\sqrt{Np(1-p)}} \tag{125}$$

follows a Normal distribution with $\mu = Np$ and $\sigma = \sqrt{Np(1-p)}$.

An approximate two-tailed 95% confidence interval for the number of successes X is then,

$$Np - 1.96 \cdot \sqrt{Np(1-p)} \leqslant X \leqslant Np + 1.96 \cdot \sqrt{Np(1-p)} \tag{126}$$

1.7.4.1 Binomial Distribution I:

What is the probability of rolling exactly 4 sixes out of 20 rolls of a fair 6-sided die?

Let's define a success as rolling a six.

$$\begin{aligned} \boldsymbol{Pr}(\mathsf{success}) &= 1/6 && (127)\\ \boldsymbol{Pr}(\mathsf{failure}) &= 1-\boldsymbol{Pr}(\mathsf{success}) = 5/6 && (128)\\ \boldsymbol{Pr}(\mathsf{X}=4) &= \binom{20}{4}\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^{20-4} = 0.2 = 20\% && (129) \end{aligned}$$

1.7.4.2 Binomial Distribution II:

What is the probability of getting a score of 12 or more on a test of 30 true-false questions?

Let's define a success as getting the question right.

$$\begin{aligned} \boldsymbol{Pr}(\mathsf{success}) &= 1/2 && (130)\\ \boldsymbol{Pr}(\mathsf{failure}) &= 1-\boldsymbol{Pr}(\mathsf{success}) = 1/2 && (131) \end{aligned}$$

Now, the problem is a bit more tedious, because we have to add up the probabilities for $\mathsf{k} = 12, 13, 14...29, 30$. Instead, we can calculate the probability of getting $\mathsf{k} = 0, 1, 2, 3...11$ questions right, and then 1 minus this probability is the probability we want. Thus,

$$\begin{aligned} \boldsymbol{Pr}(\mathsf{X} \geqslant 12) &= 1-\boldsymbol{Pr}(\mathsf{X}<12) && (132)\\ &= 1-\sum_{\mathsf{k}=0}^{11}\binom{30}{\mathsf{k}}(0.5)^{\mathsf{k}}(0.5)^{30-\mathsf{k}} && (133)\\ &= 1-0.1 = 90\% && (134) \end{aligned}$$

1.7.4.3 Normal Approximation to the Binomial:

see jupyter notebook binomial_examples.ipynb
48 CMIP5 models are discussed in the IPCC 5th Assessment Report. How many models must agree that global temperatures will increase by 2100 so that we can say with 95% certainty that the models do not agree purely by chance? What is the 95% confidence interval on the number of models with increasing temperatures under the null hypothesis?

Here, let a success be that the model says global temperatures will increase. Our null hypothesis is that the models randomly guess whether global temperatures will increase - thus, there is a 50% chance that any one model will predict a temperature increase ($\mathsf{p} = 0.5$). We want to know k^* such that:

$$\boldsymbol{Pr}(\mathsf{X} \geqslant \mathsf{k}^* | \mathsf{H}_0) \leqslant 0.05 \tag{135}$$

That is, k^* is the number of models that must show a temperature increase for us to believe it is more than chance (that the null hypothesis can be rejected).

$$\sum_{\mathsf{k}=\mathsf{k}^*}^{48} \binom{48}{\mathsf{k}} (0.5)^{\mathsf{k}} (1-0.5)^{48-\mathsf{k}} \leqslant 0.05 \tag{136}$$

This would take a long time by hand, however, let's check if the Normal approximation to the Binomial applies.

1. N is large ($\mathsf{N} = 48$) ✓
2. $\mathsf{Np} = 24 \geqslant 10$ ✓
3. $\mathsf{N}(1-\mathsf{p}) = 24 \geqslant 10$ ✓

So, we can use the Normal approximation for large N (here we choose a one-sided test):

$$\boldsymbol{Pr}\left(\mathsf{Z} > \frac{\mathsf{k}^* - 48 \cdot 0.5}{\sqrt{48 \cdot 0.5 \cdot (1-0.5)}}\right) = 0.05 \tag{137}$$

$$\frac{\mathsf{k}^* - 48 \cdot 0.5}{\sqrt{48 \cdot 0.5 \cdot (1-0.5)}} = 1.645 \tag{138}$$

$$\mathsf{k}^* \geqslant 30. \tag{139}$$

So, at least 30 models must show increasing temperatures to reject the null hypothesis that the warming is due to random chance. As expected, more than half of the models must show an increase.

The 95% confidence interval under the null hypothesis is:

$$\mathsf{Np} \pm 1.96 \cdot \sqrt{\mathsf{Np}(1-\mathsf{p})} \tag{140}$$

$$24 \pm 1.96 \cdot \sqrt{24(1-.5)} = 24 \pm 6.8 \tag{141}$$

$$17.2 \leqslant \mathsf{X} \leqslant 30.8 \tag{142}$$

1.7.5 Poisson Approximation & Rates

The Poisson distribution applies when you are counting the number of objects in a certain interval. The interval can be in space (volume, area or length) or time. You know the average number of counts per unit interval, and wish to know the chance of actually observing various numbers of objects or events. We denote the associated random variable N, since they are actual counts.

$$\mathsf{N} \Rightarrow \text{Poisson}(\lambda) \tag{143}$$

There are three necessary and sufficient conditions for a Poisson Distribution.

1. Two or more events cannot occur simultaneously. This means that the events themselves occupy negligible space (e.g. volume, area, length, time).
2. Events occur at an average rate of λ (per unit e.g. volume, area, length, time). This means that λ cannot be a function of space or time.
3. Events occur independently (i.e. they do not know about each other)

The probability mass function of a Poission is defined by the probability that $\mathsf{N} = \mathfrak{n}$ in a given interval of magnitude $\mathfrak{t}$ according to:

$$\boldsymbol{Pr}(\mathsf{N} = \mathfrak{n}) = \frac{(\lambda\mathfrak{t})^{\mathfrak{n}}}{\mathfrak{n}!} e^{-\lambda\mathfrak{t}} \qquad \lambda > 0, \mathfrak{t} > 0, \text{ and } \mathfrak{n} = 0, 1, \ldots \tag{144}$$

The first and second moments (mean and variance) are given by:

$$\text{mean } = \lambda\mathfrak{t} \qquad\qquad \text{variance } = \lambda\mathfrak{t} \tag{145}$$

Estimating $\hat{\lambda}$ from your data is quite straightforward. Let N be the number of observed events in time $\mathfrak{t}$, and assume that N is well-modelled as a Poisson Distribution with unknown rate parameter λ (where λ has units of "events per unit time"). The rather obvious formula for estimating the rate parameter is then simply the number of events divided by the time over which they were observed:

$$\hat{\lambda} = \frac{\mathsf{N}}{\mathfrak{t}} \tag{146}$$

The standard deviation (or standard error) of this estimator is

$$\sigma_{\hat{\lambda}} = \sqrt{\frac{\lambda}{\mathfrak{t}}} \tag{147}$$

Putting these together, the approximate 95% confidence interval for the true parameter λ (assuming the Central Limit Theorem applies, which it does for $\mathsf{N} > 30$ or so) is given by

$$\boldsymbol{Pr}\left(\hat{\lambda} - z_{\alpha/2}\sigma_{\hat{\lambda}} \leqslant \lambda \leqslant \hat{\lambda} + z_{\alpha/2}\sigma_{\hat{\lambda}}\right) = 0.95 \tag{148}$$

1.7.5.1 Poisson rate confidence interval

Let's say we count 137 events in 44 minutes. Our estimated rate parameter is

$$\hat{\lambda} = \frac{N}{t} = \frac{137}{44} \approx 3.11 \text{ events per minute} \tag{149}$$

The approximate standard error for the estimated rate parameter is

$$\sigma_{\hat{\lambda}} = \frac{\sqrt{N}}{t}\frac{\sqrt{137}}{44} \approx 0.27 \text{ events per minute} \tag{150}$$

and so the approximate 95% confidence interval for the true, but unknown, rate parameter is

$$\hat{\lambda} - 1.96\sigma_{\hat{\lambda}} \leqslant \lambda \leqslant \hat{\lambda} + 1.96\sigma_{\hat{\lambda}} \Rightarrow 2.58 \leqslant \lambda \leqslant 3.64 \tag{151}$$

It turns out that the Poisson Distribution has a close relationship with the Binomial Distribution. That is, for $n \to \infty$, $p \to 0$, with $np \to \lambda \neq 0$, the Binomial Distribution converges to the Poisson Distribution with parameter λ. In practice, the Binomial Distribution may be approximated by the Poisson when $p < 0.5$ and $n > 20$.

1.8 Non-parametric Tests

Thus far, most of the statistics we have used assume an underlying distribution (typically normal). However, there may be instances where you do not think that the data is normally distributed - in which case, you might want to use a non-parametric statistical test.

1.8.1 Signs Test

For example, say you want to test the null hypothesis that the means of two paired data sets x_i, y_i are the same. In this case, one could use the *Signs Test*, also known as the *Wilcoxon Test*. Instead of testing the means, this test uses the medians ($\tilde{\mu}$) of the two distributions to determine if they are equal. We formulate the hypothesis in the following way.

$$H_0 \;:\; \boldsymbol{Pr}(y_i > x_i) = 0.5 \tag{152}$$

$$H_1 \;:\; \boldsymbol{Pr}(y_i > x_i) \neq 0.5 \tag{153}$$

To do this test, you replace each pair with a signed integer in the following manner:

$$y_i > x_i \to +1 \tag{154}$$

$$y_i > x_i \to -1 \tag{155}$$

$$\tag{156}$$

The null hypothesis would suggest that there are a similar number of +1 and -1. With this setup, we now have a bunch of Bernoulli trials with a success (+1) and a failure (-1). Thus, we can use the Binomial Distribution to determine the probability of getting a certain number of +1 versus -1. Prof. Dennis Hartmann gives a nice example in Chapter 1.

1.8.2 Runs Test (Wald-Wolfowitz runs test)

The runs test is a non-parametric test to check whether a data set is random or not. For example (taken from Wikipedia), imagine a time series of anomalies as shown below, where "+" denotes a positive anomaly and "-" denotes a negative anomaly:

$$++++--\,-++\,--+++++\,----\tag{157}$$

We now separate this series into "runs"...specifically, there are $\mathsf{R} = 6$ runs, 3 of which consist of "+" and the others of "-". The runs test uses a null hypothesis that the data is random. Under this null hypothesis, the number of runs (R) in a sequence of N elements is a random variable whose conditional distribution given the observation of N_+ positive values and N_- negative values ($\mathsf{N} = \mathsf{N}_+ + \mathsf{N}_-$) has the following properties:

$$\mu = 1 + \frac{2\mathsf{N}_+\mathsf{N}_-}{\mathsf{N}} \tag{158}$$

$$\sigma^2 = \frac{2\mathsf{N}_+\mathsf{N}_-(2\mathsf{N}_+\mathsf{N}_- - \mathsf{N})}{\mathsf{N}^2(\mathsf{N}-1)} = \frac{(\mu-1)(\mu-2)}{\mathsf{N}-1} \tag{159}$$

If N_+ and N_- are sufficiently large (say, greater than 30) then the number of runs R is well modeled by a Normal distribution with parameters μ and σ given above.

Example: NOTEBOOK RUNSTEST.IPYNB

1.8.3 Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test (or "KS test") tests the equality of two continuous, one-dimensional probability distributions. The most standard version tests whether a particular sample distribution is the same as a specific reference distribution.

A few important notes about the ks-test:

- this is a non-parametric test, so you do not need to know what the true distribution of your data is or test against some null hypothesis of a particular distribution
- the test is sensitive to both *location* and *shape* and cannot tell you why the distributions are different (e.g. is the sample shifted compared to the reference? is the sample distribution wider than the reference?)

The test works by comparing the CDFs (cumulative density functions) of the sample of length N and reference distributions. Specifically, the difference between the two CDFs is computed, and the *maximum difference*, denoted as D, is used as the test statistic. The null hypothesis is rejected at the significance level α if

$$\sqrt{\mathsf{N}}\mathsf{D} > \mathsf{K}_\alpha \tag{160}$$

where K_α is defined as

$$\boldsymbol{Pr}(\mathsf{K} \leqslant \mathsf{K}_\alpha) = 1 - \alpha \tag{161}$$

and the probability density function of K is defined as

$$\boldsymbol{Pr}(\mathsf{K} \leqslant \mathsf{x}) = 1 - 2\sum_{\mathsf{j}=1}^{\infty}(-1)^{\mathsf{j}-1}\boldsymbol{e}^{-2\mathsf{j}^2\mathsf{x}^2} \tag{162}$$

A few additional points

- If you are specifically interested in whether your sample distribution is normal, other tests may be better suited (e.g. the Shapiro-Wilks or the Anderson-Darling test)
- Do not estimate the parameters of the *reference distribution* from the data. The test is not valid if you do. Thus, if you are comparing to a normal distribution and don't know the true mean and standard deviation of your sample population, you should standardize your data first and compare the standardized sample to the standard normal.

Finally, note that the above discussion only applies when you wish to compare a single sample to some reference distribution. What if instead you wish to compare two sample distributions? For that, you can use the Two Sample KS-Test.

Example: RUN NOTEBOOK KS_TEST.IPYNB