# Class Notes: ATM 552 Objective Analysis

## 1. Review of Basic Statistics

We will review a few features of statistics that come up frequently in objective analysis of geophysical, medical and social data. The emphasis here is not on mathematical sophistication, but in developing an ability to use relatively common statistical tests correctly.

### 1.1 Some Fundamental Statistical Quantities

**The Mean:**

The sample mean of a set of values, $x_i$, is given by

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1.1}$$

This estimate of the true mean μ is unbiased. The sample mean, or average, is an unbiased estimate of the mean. The mean is the first moment about zero. The mean is to be distinguished from the median, which is the value in the center of the population (or the average of the two middle values, if the sample contains an even number of examples), and the mode, which is the most frequently occurring value.

**The Variance:**

The sample variance of a set of values is given by

$$\overline{x'^2} = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2 \tag{1.2}$$

The division by N-1 instead of the expected N is to obtain an unbiased estimate of the variance. To see why this is so check any standard textbook on mathematical statistics, where you will find a fairly long derivation that I don't want to reproduce here. Basically, the variance is biased low because the sample mean is uncertain and its uncertainty gives the sample variance a low bias. Using N-1 as the sample size corrects for that. The variance is the second moment about the mean. It is more efficient (because it can be done with one loop and roundoff error is reduced) to compute the variance using the following relation:

$$\overline{x'^2} = \frac{N}{N-1}\left(\overline{x^2} - \overline{x}^2\right) \tag{1.3}$$

Where, of course:

$$\overline{x^2} = \frac{1}{N}\sum_{i=1}^{N} x_i^2$$

**The Standard Deviation:**

The standard deviation is the square root of the variance. We often denote it with the symbol $\sigma$, the sample standard deviation is $s$.

$$s = \sqrt{\overline{x'^2}} \tag{1.4}$$

**Higher Moments:**

We can define an arbitrary moment about the mean as:

$$m_r = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^r \tag{1.5}$$

So that $m_2$ is the variance, $m_3$ is the skewness, and $m_4$ is the kurtosis. These can be non-dimensionalized by defining

$$a_r = \frac{m_r}{\sigma^r} \tag{1.6}$$

Where $\sigma$ is again the standard deviation, or the square root of the second moment about the mean.

The moment coefficient of skewness, $a_3$, indicates the degree of asymmetry of the distribution about the mean. If $a_3$ is positive then the distribution has a longer tail on the positive side of the mean, and vice versa. The coefficient of skewness for a Normal distribution is zero.

The moment coefficient of kurtosis (Greek word for bulging), $a_4$, indicates the degree to which the distribution is spread about the mean value. The fourth moment about the mean is the Pearson measure of kurtosis and tends to measure the thickness of the tails of the distribution. The Normal distribution is in the region called *mesokurtic* and has a coefficient of kurtosis of 3. Distributions with very flat distributions near the mean, with high coefficients of kurtosis, are called *platykurtic* (Greek *platys*, meaning broad or flat). Distributions that are strongly peaked near the mean have low coefficients of kurtosis and are called *leptokurtic* (Greek for *leptos*, meaning small or narrow). In many statistics packages the coefficient of kurtosis has the value for a normal distribution, 3, subtracted from it, so that platykurtic distributions have negative coefficients and vice versa.

## 1.2 Probability Concepts and Laws

One view of probability is the frequency view. If you have some large number of opportunities for an event to occur, then the number of times that event actually occurs, divided by the number of opportunities for it to occur is the probability. The probability varies between zero and one. The frequentist view has a solid foundation in the Weak Law of Large Numbers which states that if you have random number between zero and one, the sum of this number divided by the sample size approaches the probability with arbitrary precision for large sample size. Another more subjective view attributed to Rev. Thomas Bayes (1701-1761) figures that in many cases one is unlikely to have a large sample with which to measure the frequency of occurrence, and so one must take a more liberal and subjective view. Bayesian inference is given that name for its frequent use of Bayes Theorem, which it uses to take into account a priori information, that may not be derivable from a frequentist point of view.

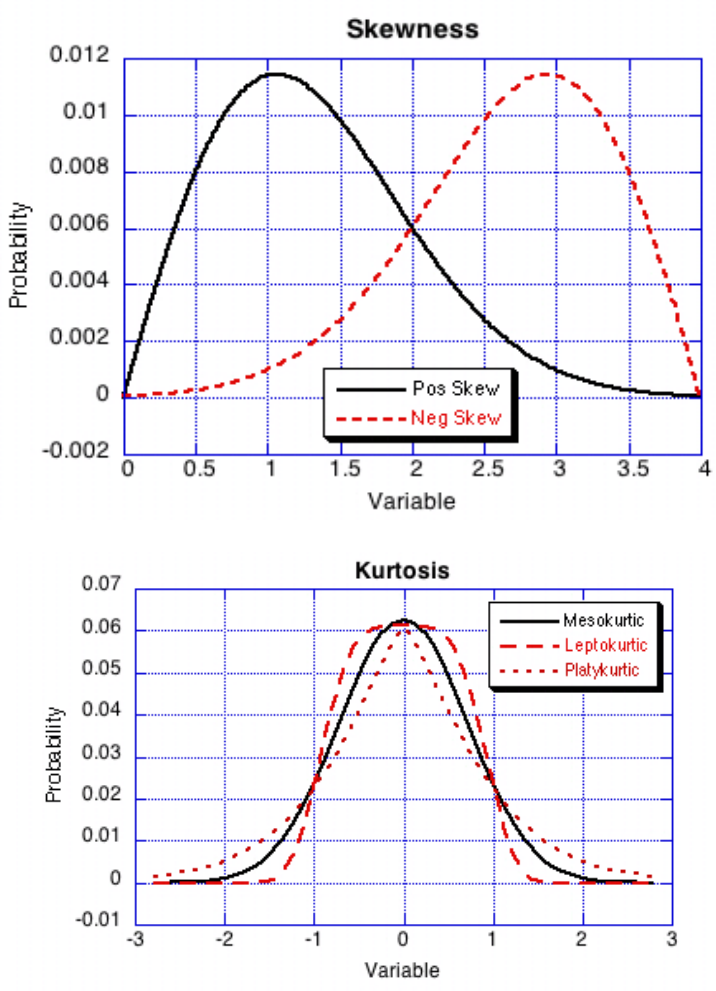



Fig. 1.1 Illustrations of skewness and kurtosis of probability distributions .

A nice philosophic discussion of the use of statistics in science from the Bayesian point of view can be found in Howson and Urbach(2006), which is available in paperback for modest cost. A more mathematical presentation is given by Jaynes(2003).

*Unions and Intersections of Probability – Venn Diagram*

The probability of some event $E$ happening is written as $P(E)$. The probability of $E$ not happening must be 1 - $P(E)$. The probability that either or both of two events $E_1$ and $E_2$ will occur is called the union of the two probabilities and is given by,

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) \tag{1.7}$$

where $P(E_1 \cap E_2)$ is the probability that both events will occur, and is called the intersection. It is the overlap between the two probabilities and must be subtracted from the sum. This is easily seen using a Venn diagram, as below. 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.

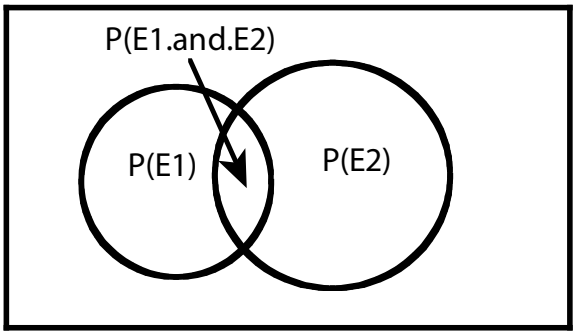



*Fig. 1.2 Venn Diagram illustrating the intersection of two probabilities.*

Another important concept is conditional probability. We write the probability that $E_2$ will occur, given that $E_1$ has occurred as the postulate,

$$P(E_2 \mid E_1) = \frac{P(E_1 \cap E_2)}{P(E_1)} \tag{1.8}$$

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

$$P(E_1 \cap E_2) = P(E_2 \mid E_1) \bullet P(E_1) = P(E_1 \mid E_2) \bullet P(E_2) \tag{1.9}$$

If the two events are completely independent such that $P(E_1 \mid E_2) = P(E_1)$, then we get,

$$P(E_1 \cap E_2) = P(E_1) \bullet P(E_2) \qquad (1.10)$$

This is the definition of statistical independence.

For example, if the probability of getting heads on a coin flip is 0.5, and one coin flip is independent every other one, then the probability of getting heads (or tails) N times in a row is $(0.5)^N$.

**Bayes Theorem:**

Bayes Theorem:

Let $E_i$, $i=1,2,3, \ldots n$ be a set of $n$ events, each with positive probability, that partition a set $S$, in such a way that.

$$\bigcup_{i=1}^{n} E_i = S \qquad \text{and } E_i \cap E_j = \varnothing \text{ for } i \neq j$$

This means the events include all the possibilities in S and the events are mutually exclusive.
For any event $B$, also defined on $S$, with positive probability $P(B) > 0$, then,

$$P\left(E_j \mid B\right) = \frac{P(B \mid E_j)\, P(E_j)}{\sum_{i=1}^{n} P(B \mid E_i) P(E_i)} \qquad (1.11)$$

Bayes Theorem can be derived from the following definitions and postulates.

Define the conditional probability of an event E, given that an event B has occurred.

$P\left(E \mid B\right) = \frac{P(E \cap B)}{P(B)}$, which can be rearranged as, $P(E \cap B) = P\left(E \mid B\right) \bullet P(B)$, which is far more intuitive.

We also know that, $P(B) = \sum_{i=1}^{n} P\left(B \mid E_i\right) \bullet P(E_i)$, which must be true of the sum, if the $E_i$ are all the possible outcomes. Can you see why each of these statements must be true? Can you use them to derive Bayes Theorem?

***Bayes Theorem Example:***

Bayes Theorem has been useful in the medical field. Consider this example. You have a sensitive test for a disease, which tells whether the patient has the disease or not. The probability that any individual has the disease is one in one thousand. The test never gives a false negative result by saying that the patient does not have the disease, when in fact the patient does. The chance of a false positive is also small, say 5%. If you have the test and it is positive, what is the probability that you actually have the disease? Most people answer 95% to this question, since only a 5% chance exists that the test has given a false positive. Let's use Bayes theorem, though, to get the real answer. We have P(D)=0.001. We have P(-|D)=0.0, the probability that you get a negative diagnosis when you have the disease (D), is zero. We have P(+|notD)=0.05. Let's use Bayes theorem (1.11) to compute P(D|+),the probability that we have the disease, given a positive test.

$$P(D\mid +)=\frac{P(+\mid D)\bullet P(D)}{P(+\mid D)\bullet P(D)+P(+\mid notD)\bullet P(notD)} \tag{1.12}$$

To solve the problem you have to realize that, if a person has the disease, then they get either a positive or negative result on the test so that $P(+\mid D)+P(-\mid D)=1$. Therefore, $P(+\mid D)=1-P(-\mid D)=1$. Plugging in numbers gives,

$$P(D\mid +)=\frac{1.0*0.001}{1.0\bullet 0.001+0.05\bullet 0.999}=0.0196=\frac{1}{51}$$

So, whereas in a kneejerk answer you might say the chance that you have the disease is 95%, the actual probability is only one in 51! The failure of the quick reaction is to not take into account the probability of this disease in the whole population, or the total probability of having the disease. If you forget about the formulas and think about numbers it is easy to see why this is so. We have been told that one person in a thousand has the disease. Also, because of the false positive rate of 5%, if 1000 people take the test 50 will get false positives on the test. The one person with the disease also gets a positive, as the test never gives false negatives. So of the 51 people who get positive test results, only one actually has the disease. Of course, this result is true only if you average over an infinite ensemble of groups of 1000, and one should not expect this result every time. The actual answers that statistics can give must be phrased in terms of intervals at some probability level.

Here's another example in the same vein, slightly more complicated. You have a large sample of the results of a particular test for breast cancer. C is the event that the patient has cancer, and $C^c$ is the alternative that the patient does not have cancer. The probability of a patient having cancer is $P(C) = 0.001$. Event B is a positive biopsy, indicating cancer, $B^c$ is its alternative. The probability of a positive biopsy, when the patient actually has cancer is $P(B|C) = 0.9$. The probability of a false positive from a biopsy is $P(B|C^c) = 0.01$, so the test gives a false positive in one out of one hundred tries.

The question is, what is the probability that the patient actually has cancer, if the test gives a positive result? So we want to know *P(C|B)*. According to Bayes Theorem, the answer is:

$$P(C \mid B) = \frac{P(B \mid C) \bullet P(C)}{P(B \mid C) \bullet P(C) + P\left(B \mid C^c\right) \bullet P(C^c)} \tag{1.13}$$

Plugging in the numbers above gives an answer of P(C|B) = 0.089. That is, if you test positive for cancer, under the conditions of this problem, then the chance that you actually have cancer is only 8.9%, less than 1 in 10. Tracking the reasons for this back through Bayes Theorem to the numbers shows that the high probability of a false positive (91%) arises because of the low probability of the occurrence of cancer, and small false positive rate of 1%. Bayes theorem might also suggest why extremely rare diseases are often not well diagnosed and treated.

What relevance do these examples have to what we do? Suppose we have a theory *T*, which we wish do test with data *D*. Formally, we want to evaluate the probability that the theory is true, given the data available. Let's look at this with Bayes theorem.

$$P(T \mid D) = \frac{P(D \mid T) \bullet P(T)}{P(D \mid T) \bullet P(T) + P(D \mid T^c) \bullet P(T^c)} \tag{1.14}$$

Here $T^c$ is the null hypothesis and *T* is the theory that we would like to say is true. The statistical tests that we often apply ask, $P\left(D \mid T^c\right)$, or "What is the probability of seeing this data result, *D*, given our null hypothesis $T^c$ ?" To evaluate the probability that our theory is true, given our data *P(T|D)*, by (1.14) we need to know the probability that our theory is true, which we do not. One can see from the previous examples, that the conclusions based on any test can be radically altered by the total probability. If the probability that we see the data we do, given the null hypothesis, is only 5%, but the probability of our theory being correct is only one in 100, what is the probability that our theory is correct given that our data can reject the null hypothesis with 95% confidence? If you have chosen a particularly unlikely theory, then seemingly strong data support for it may not make the actual probability that it is true very large. How do you determine the likelihood that a theory is true, *P(T)*? This line of reasoning exposes a flaw that Bayesian's see in the frequentist view of probability. It also suggests why people can often find support in data for bizarre theories. In doing statistical significance tests we often think that $P(T) = 1 - P\left(D \mid T^c\right)$, but this is clearly not valid if (1.14) is true. That is, we assume that the probability that the hypothesis is true, is equal to one minus the probability that we would see the data, given the null hypothesis. We have therefore, made prior assumptions that we have not considered very carefully. If you were trained as a frequentist, you may find this section disturbing. Often people talk about 'assessing

your priors'. In your evaluation of probability, you unconsciously factor in your 'pre-conceived notions'. You should consciously consider accounting for this.

### 1.3 Probability Distributions

The probability that a randomly selected value of a variable $x$ falls between the limits $a$ and $b$ can be written:

$$P(a \le x \le b) = \int_a^b f(x)\ dx \tag{1.15}$$

This expression defines the probability density function $f(x)$ in the continuous case. Note that the probability that the variable $x$ will assume some particular value, say $c$, is exactly zero. $f(x)$ is not the probability that $x$ will assume the value $x_1$. To obtain a probability one must integrate the probability density function between distinct limits. The probability density must have the following characteristics:

$f(x) \ge 0$ for all $x$ within the domain of $f$

$$\int_{-\infty}^{\infty} f(x)dx = 1 \tag{1.16}$$

The moments of the distribution can be obtained from the probability density using the following formula

$$m_r = \int_{-\infty}^{\infty} (x-\mu)^r\ f(x)dx \tag{1.17}$$

where $\mu$ is the true mean, and so the moments are taken about the mean.

The cumulative distribution function $F(x)$ can be defined as the probability that a variable assumes a value less than $x$:

$$F(x) = \int_{-\infty}^{x} f(t)dt \tag{1.18}$$

It immediately follows that

$$P(a \le x \le b) = F(b) - F(a)$$
$$\frac{dF}{dx} = f(x) \tag{1.19}$$

## 1.4 The Normal Distribution:

The Normal distribution is one of the most important in nature. Most observables are distributed normally about their means, or can be transformed in such a way that they become normally distributed. It is important to verify in some way that your random variable is Normally distributed before using Gaussian-Normal statistical tests, however. We can assume that we have a standardized random variable $z$ derived from some unstandardized random variable $x$:

$$z = \frac{(x - \mu)}{\sigma} \tag{1.20}$$

So standardized means zero mean and unit variance. Then, if $z$ (and $x$) is normally distributed, the cumulative distribution function is:

$$F(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}t^2\right\} dt \tag{1.21}$$

With the probability density function given by the part inside the integral, of course.

If we use the unstandardized random variable $x$, then the form of the probability density is:

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left[\frac{x - \mu}{\sigma}\right]^2\right\} \tag{1.22}$$

In this formula μ and σ are actually the mean and standard deviation. Of course, the probability density is only defined relative to the cumulative distribution, and this explains why the $\sigma$ appears in the denominator of the constant expression multiplying the exponential immediately above. It arises when the transformation is made in the variable of integration.

Table of the Cumulative Normal Probability distribution F(z).

| *z* | **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 |
| **2.5** | 0.9938 | 0.9940 | 0.9941 | 0.9943 | 0.9945 | 0.9946 | 0.9948 | 0.9949 | 0.9951 | 0.9952 |
| **2.6** | 0.9953 | 0.9955 | 0.9956 | 0.9957 | 0.9959 | 0.9960 | 0.9961 | 0.9962 | 0.9963 | 0.9964 |
| **2.7** | 0.9965 | 0.9966 | 0.9967 | 0.9968 | 0.9969 | 0.9970 | 0.9971 | 0.9972 | 0.9973 | 0.9974 |
| **2.8** | 0.9974 | 0.9975 | 0.9976 | 0.9977 | 0.9977 | 0.9978 | 0.9979 | 0.9979 | 0.9980 | 0.9981 |
| **2.9** | 0.9981 | 0.9982 | 0.9982 | 0.9983 | 0.9984 | 0.9984 | 0.9985 | 0.9985 | 0.9986 | 0.9986 |
| **3.0** | 0.9987 | 0.9987 | 0.9987 | 0.9988 | 0.9988 | 0.9989 | 0.9989 | 0.9989 | 0.9990 | 0.9990 |
| **3.1** | 0.9990 | 0.9991 | 0.9991 | 0.9991 | 0.9992 | 0.9992 | 0.9992 | 0.9992 | 0.9993 | 0.9993 |
| **3.2** | 0.9993 | 0.9993 | 0.9994 | 0.9994 | 0.9994 | 0.9994 | 0.9994 | 0.9995 | 0.9995 | 0.9995 |
| **3.3** | 0.9995 | 0.9995 | 0.9995 | 0.9996 | 0.9996 | 0.9996 | 0.9996 | 0.9996 | 0.9996 | 0.9997 |
| **3.4** | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9998 |

The probability that a normally distributed variable falls within one standard deviation of its mean value is given by:

$$P(-1 \le z \le 1) = \int_{-1}^{+1} f(z)\,dz = 68.27\% \tag{1.23}$$

and similarly for 2 and 3 standard deviations:

$$P(-2 \le z \le 2) = \int_{-2}^{+2} f(z)\,dz = 95.45\% \tag{1.24}$$

$$P(-3 \le z \le 3) = \int_{-3}^{+3} f(z)\,dz = 99.73\% \tag{1.25}$$

Thus there is only a 4.55% probability that a normally distributed variable will fall more than 2 standard deviations away from its mean. This is the two-tailed probability. The probability that a normal variable will exceed its mean by more then $2\sigma$ is only half of that, 2.275%, since the Normal distribution is symmetric.

**1.5 Testing for Significance using the Normal Distribution:**

As it turns out, many geophysical variables are approximately normally distributed. This means that we can use the theoretical normal probability distribution to calculate the probability that two means are different, etc. Unfortunately, to do this we need to know the true mean $\mu$ and the true standard deviation $\sigma$, *a priori*. The best that we are likely to have are the sample mean $\overline{x}$ and the sample standard deviation *s* based on some sample of finite size *N*. If *N* is large enough we can use these estimates to compute the *z* statistic. Otherwise we need to use the Student *t* statistic, which is more appropriate for small samples. In geophysical applications we can usually assume that we are sampling from an infinite population.

For an infinite population the standard deviation of the sampling distribution of means is given by:

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{N}} = \text{ the standard error of estimate of the mean.}$$

Here $\sigma$ is the standard deviation of the population and *N* is the number of data used to compute the sample mean. So as you average together observations from a population

of standard deviation $\sigma$, the standard deviation of those averages goes down as the inverse of the square root of the sample size N. The standard variable used to compare a sample mean to the true mean is thus:

$$z=\frac{\overline{x}-\mu}{\sigma_{\overline{x}}}=\frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{N}}} \tag{1.26}$$

The statistic $z$ is thus the number of standard errors that the sample mean deviates from the true mean, or the null hypothesis mean. If the variable is normally distributed about its mean, then $z$ can be converted into a probability statement. This formula needs to be altered only slightly to provide a significance test for differences between means:

$$z=\frac{\overline{x}_1-\overline{x}_2-\Delta_{1,2}}{\sqrt{\frac{\sigma_1^2}{N_1}+\frac{\sigma_2^2}{N_2}}} \tag{1.27}$$

Here the sample sizes for computing the two means and the two standard deviations are different. $\Delta_{1,2}$ is the expected difference between the two means, which is often zero in practice.

**Small Sampling Theory:**

When the sample size is smaller than about 30 we cannot use the $z$ statistic, above, but must use the Student's $t$ distribution; or when comparing variances, the chi-squared distribution. Since the Student's $t$ distribution approaches the normal distribution for large $N$, there is no theoretical reason to use the normal distribution in preference to Student's $t$, although it may be more convenient to do so sometimes.

The Student's $t$ distribution is derived in exact analogy with the $z$ statistic:

$$t=\frac{\overline{x}-\mu}{\frac{s}{\sqrt{N-1}}}=\frac{\overline{x}-\mu}{\frac{\hat{s}}{\sqrt{N}}};\ \hat{s}=\sqrt{\frac{N}{N-1}}\,s \tag{1.28}$$

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

$$f(t)=\frac{f_o(v)}{\left(1+\frac{t^2}{v}\right)^{\frac{(v+1)}{2}}} \tag{1.29}$$

Where $\nu = N - 1$ is the number of degrees of freedom and $f_O(\nu)$ is a constant that depends on $\nu$ and makes the area under the curve $f(t)$ equal to unity.

Unlike the $z$ distribution, the $t$ distribution depends on the size of the sample. The tails of the distribution are longer for smaller degrees of freedom. For a large number of degrees of freedom the $t$ distribution approaches the $z$ or normal distribution. Note that, although we sometimes speak of the $t$ distribution and contrast it with the normal distribution, the $t$ distribution is merely the probability density you expect to get when you take a small sample from a normally distributed population. The Student 's $t$ distribution is the most commonly used in means testing, and perhaps in all of applied statistics, although non-parametric methods are becoming more standard nowadays.

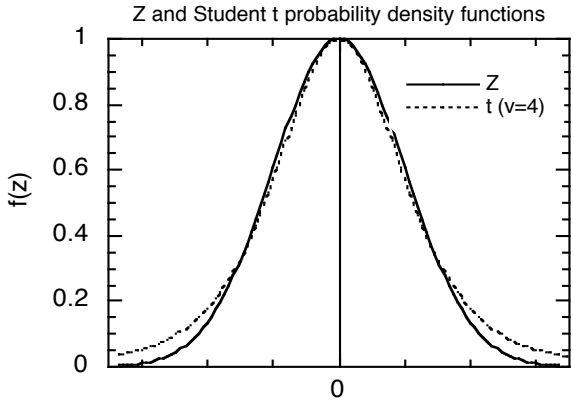



*Fig. 1.3 PDF of Z and Student-t with four dof.*

**Confidence intervals:**

Values of the t statistic and the $z$ statistic for specified probability levels and degrees of freedom are given in tables. In such tables, $t_{0.025}$ is the value of $t$ for which only 0.025, 2.5%, of the values of $t$ would be expected to be greater (right-hand tail). $t_{-0.025} = -t_{0.025}$ is the value of $t$ for which only 2.5% of the values of $t$ obtained from a normally distributed sample would be expected to be less. Note that the $t$ distribution is symmetric. The values of $t$ are the integrals under the probability density function as shown below.

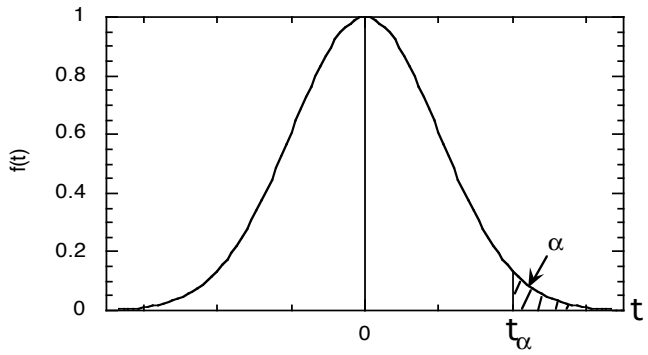



*Fig. 1.4 Illustration of relation of t-statistic pdf to probability measure alpha.*

There is a 95% probability that any sampled $t$ statistic falls in the interval

$$t_{-.025} < \frac{\overline{x} - \mu}{s} \cdot \sqrt{N-1} < t_{.025} \tag{1.30}$$

From this we can deduce that the true mean $\mu$ is expected with 95% confidence to lie in the interval:

$$\overline{x} - t_{.025} \cdot \frac{s}{\sqrt{N-1}} < \mu < \overline{x} + t_{.025} \cdot \frac{s}{\sqrt{N-1}} \tag{1.31}$$

In general, confidence limits for population means can be represented by

$$\mu = \overline{x} \pm t_c \cdot \frac{s}{\sqrt{N-1}} \tag{1.32}$$

Where $t_c$ is the critical value of the $t$ statistic, which depends on the number of degrees of freedom and the statistical confidence level desired. Comparing this with the confidence limits derived using the $z$ statistic, which is only appropriate for large samples where the standard deviation can be assumed known:

$$\mu = \overline{x} \pm z_c \cdot \frac{\sigma}{\sqrt{N}} \tag{1.33}$$

we see that the small sample theory replaces the $z$ statistic with the $t$ statistic and the standard deviation by a modified sample standard deviation:

$$\hat{s} = s\sqrt{\frac{N}{N-1}} \tag{1.34}$$

**Differences of Means:**

Suppose two samples of size $N_1$ and $N_2$ are drawn from a normal population whose standard deviations are equal. Suppose the sample means are given by $x_1$ and $x_2$ and the sample standard deviations are $s_1$ and $s_2$. To test the null hypothesis $H_o$ that the samples come from the same population ($\mu_1=\mu_2$ as well as $\sigma_1=\sigma_2$) use the $t$ score given by:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}; \text{ where } \sigma = \sqrt{\frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2 - 2}} \tag{1.35}$$

$$\text{and } v = N_1 + N_2 - 2$$

**Chi-Squared Distribution: Tests of Variance**

Sometimes we want to test whether sample variances are truly different. For this we can define the Chi-Squared Statistic. Define:

$$\chi^2 = (N-1)\frac{s^2}{\sigma^2} \tag{1.36}$$

Draw $\chi^2$ from a normal distribution with standard deviation $\sigma$. The samples are distributed according to:

$$f(\chi) = f_o \chi^{v-2} e^{-\frac{1}{2}\chi^2} \qquad v = N-1 \tag{1.37}$$

Note that the Chi-squared distribution is not symmetric, so that we write the 95% confidence limits as:

$$\frac{(N-1)s^2}{\chi^2_{0.025}} < \sigma^2 < \frac{(N-1)s^2}{\chi^2_{0.975}} \tag{1.39}$$

**Degrees of Freedom:**

The number of degrees of freedom is the number of independent samples *N* minus the number of parameters in the statistic that is being estimated. For example in the *t* statistic,

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{N-1}}} = \frac{\bar{x} - \mu}{\frac{\hat{s}}{\sqrt{N}}} \quad ; \quad \hat{s} = \sqrt{\frac{N}{N-1}}s \tag{1.40}$$

we calculate the sample mean *x* and the sample standard deviation *s* from the data, but the true mean must be estimated, thus $v = N - 1$. Similarly in the Chi-squared statistic,

$$\chi^2 = \frac{(N-1)\,s^2}{\sigma^2} \tag{1.41}$$

we know the sample variance $s^2$ and the sample size *N*, but we must estimate the true

variance so that $\nu = N - 1$. Some would argue, however, that we need the sample mean to estimate the sample variance, so that in fact $\nu = N - 2$, but this is heresy. If it makes a significant difference, your sample is too small anyway.

**Note:** It is assumed here that the $N$ samples are **independent samples**. Often $N$ observations of a geophysical variable are not independent and we must try to estimate the number of independent observations in the sample. For example, the geopotential height is highly auto-correlated so that each day's value is not independent from the previous or following day's. You can't improve the your ability to know a 5-day wave by sampling every 3 hours instead of every 6, for example. This is discussed further in sections 1.9 and 6.1.5.

**F Statistic**

Another statistic that we will find useful in testing power spectra is the F-statistic. If $s_1^2$ and $s_2^2$ are the variances of independent random samples of size $N_1$ and $N_2$, taken from two Normal populations having the same variance, then

$$F = \frac{s_1^2}{s_2^2} \tag{1.42}$$

is a value of a random variable having the F distribution with the parameters $\nu_1 = N_1 - 1$ and $\nu_2 = N_2 - 1$. This statistic will be very useful in testing the significance of peaks in frequency spectra. The two parameters are the degrees of freedom for the sample variance in the numerator, $\nu_1$, and in the denominator, $\nu_2$.

Table of the t-statistic critical values for one-tailed test with ν degrees of freedom.

| ν | 0.2 | 0.1 | 0.05 | 0.025 | 0.01 | 0.005 | 0.001 | 0.0005 | 0.0001 |
|---|---|---|---|---|---|---|---|---|---|
| **1** | 1.3764 | 3.0777 | 6.3137 | 12.706 | 31.821 | 63.656 | 318.29 | 636.6 | 3185.3 |
| **2** | 1.0607 | 1.8856 | 2.9200 | 4.3027 | 6.9645 | 9.9250 | 22.328 | 31.600 | 70.706 |
| **3** | 0.9785 | 1.6377 | 2.3534 | 3.1824 | 4.5407 | 5.8408 | 10.214 | 12.924 | 22.203 |
| **4** | 0.9410 | 1.5332 | 2.1318 | 2.7765 | 3.7469 | 4.6041 | 7.1729 | 8.6101 | 13.039 |
| **5** | 0.9195 | 1.4759 | 2.0150 | 2.5706 | 3.3649 | 4.0321 | 5.8935 | 6.8685 | 9.6764 |
| **6** | 0.9057 | 1.4398 | 1.9432 | 2.4469 | 3.1427 | 3.7074 | 5.2075 | 5.9587 | 8.0233 |
| **7** | 0.8960 | 1.4149 | 1.8946 | 2.3646 | 2.9979 | 3.4995 | 4.7853 | 5.4081 | 7.0641 |
| **8** | 0.8889 | 1.3968 | 1.8595 | 2.3060 | 2.8965 | 3.3554 | 4.5008 | 5.0414 | 6.4424 |
| **9** | 0.8834 | 1.3830 | 1.8331 | 2.2622 | 2.8214 | 3.2498 | 4.2969 | 4.7809 | 6.0094 |
| **10** | 0.8791 | 1.3722 | 1.8125 | 2.2281 | 2.7638 | 3.1693 | 4.1437 | 4.5868 | 5.6939 |
| **11** | 0.8755 | 1.3634 | 1.7959 | 2.2010 | 2.7181 | 3.1058 | 4.0248 | 4.4369 | 5.4529 |
| **12** | 0.8726 | 1.3562 | 1.7823 | 2.1788 | 2.6810 | 3.0545 | 3.9296 | 4.3178 | 5.2631 |
| **13** | 0.8702 | 1.3502 | 1.7709 | 2.1604 | 2.6503 | 3.0123 | 3.8520 | 4.2209 | 5.1106 |
| **14** | 0.8681 | 1.3450 | 1.7613 | 2.1448 | 2.6245 | 2.9768 | 3.7874 | 4.1403 | 4.9849 |
| **15** | 0.8662 | 1.3406 | 1.7531 | 2.1315 | 2.6025 | 2.9467 | 3.7329 | 4.0728 | 4.8801 |
| **16** | 0.8647 | 1.3368 | 1.7459 | 2.1199 | 2.5835 | 2.9208 | 3.6861 | 4.0149 | 4.7905 |
| **17** | 0.8633 | 1.3334 | 1.7396 | 2.1098 | 2.5669 | 2.8982 | 3.6458 | 3.9651 | 4.7148 |
| **18** | 0.8620 | 1.3304 | 1.7341 | 2.1009 | 2.5524 | 2.8784 | 3.6105 | 3.9217 | 4.6485 |
| **19** | 0.8610 | 1.3277 | 1.7291 | 2.0930 | 2.5395 | 2.8609 | 3.5793 | 3.8833 | 4.5903 |
| **20** | 0.8600 | 1.3253 | 1.7247 | 2.0860 | 2.5280 | 2.8453 | 3.5518 | 3.8496 | 4.5390 |
| **21** | 0.8591 | 1.3232 | 1.7207 | 2.0796 | 2.5176 | 2.8314 | 3.5271 | 3.8193 | 4.4925 |
| **22** | 0.8583 | 1.3212 | 1.7171 | 2.0739 | 2.5083 | 2.8188 | 3.5050 | 3.7922 | 4.4517 |
| **23** | 0.8575 | 1.3195 | 1.7139 | 2.0687 | 2.4999 | 2.8073 | 3.4850 | 3.7676 | 4.4156 |
| **24** | 0.8569 | 1.3178 | 1.7109 | 2.0639 | 2.4922 | 2.7970 | 3.4668 | 3.7454 | 4.3819 |
| **25** | 0.8562 | 1.3163 | 1.7081 | 2.0595 | 2.4851 | 2.7874 | 3.4502 | 3.7251 | 4.3516 |
| **26** | 0.8557 | 1.3150 | 1.7056 | 2.0555 | 2.4786 | 2.7787 | 3.4350 | 3.7067 | 4.3237 |
| **27** | 0.8551 | 1.3137 | 1.7033 | 2.0518 | 2.4727 | 2.7707 | 3.4210 | 3.6895 | 4.2992 |
| **28** | 0.8546 | 1.3125 | 1.7011 | 2.0484 | 2.4671 | 2.7633 | 3.4082 | 3.6739 | 4.2759 |
| **29** | 0.8542 | 1.3114 | 1.6991 | 2.0452 | 2.4620 | 2.7564 | 3.3963 | 3.6595 | 4.2538 |
| **30** | 0.8538 | 1.3104 | 1.6973 | 2.0423 | 2.4573 | 2.7500 | 3.3852 | 3.6460 | 4.2340 |
| **40** | 0.8507 | 1.3031 | 1.6839 | 2.0211 | 2.4233 | 2.7045 | 3.3069 | 3.5510 | 4.0943 |
| **50** | 0.8489 | 1.2987 | 1.6759 | 2.0086 | 2.4033 | 2.6778 | 3.2614 | 3.4960 | 4.0140 |
| **75** | 0.8464 | 1.2929 | 1.6654 | 1.9921 | 2.3771 | 2.6430 | 3.2024 | 3.4249 | 3.9116 |
| **100** | 0.8452 | 1.2901 | 1.6602 | 1.9840 | 2.3642 | 2.6259 | 3.1738 | 3.3905 | 3.8615 |
| **∞** | 0.8416 | 1.2816 | 1.6449 | 1.9600 | 2.3264 | 2.5758 | 3.0902 | 3.2905 | 3.7189 |

Table of the Chi-Squared Distribution.

| ν | 0.995 | 0.990 | 0.975 | 0.950 | 0.900 | 0.100 | 0.050 | 0.025 | 0.010 | 0.005 |
|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 0.000 | 0.000 | 0.001 | 0.004 | 0.016 | 2.706 | 3.841 | 5.024 | 6.635 | 7.879 |
| **2** | 0.010 | 0.020 | 0.051 | 0.103 | 0.211 | 4.605 | 5.991 | 7.378 | 9.210 | 10.597 |
| **3** | 0.072 | 0.115 | 0.216 | 0.352 | 0.584 | 6.251 | 7.815 | 9.348 | 11.345 | 12.838 |
| **4** | 0.207 | 0.297 | 0.484 | 0.711 | 1.064 | 7.779 | 9.488 | 11.143 | 13.277 | 14.860 |
| **5** | 0.412 | 0.554 | 0.831 | 1.145 | 1.610 | 9.236 | 11.070 | 12.832 | 15.086 | 16.750 |
| **6** | 0.676 | 0.872 | 1.237 | 1.635 | 2.204 | 10.645 | 12.592 | 14.449 | 16.812 | 18.548 |
| **7** | 0.989 | 1.239 | 1.690 | 2.167 | 2.833 | 12.017 | 14.067 | 16.013 | 18.475 | 20.278 |
| **8** | 1.344 | 1.647 | 2.180 | 2.733 | 3.490 | 13.362 | 15.507 | 17.535 | 20.090 | 21.955 |
| **9** | 1.735 | 2.088 | 2.700 | 3.325 | 4.168 | 14.684 | 16.919 | 19.023 | 21.666 | 23.589 |
| **10** | 2.156 | 2.558 | 3.247 | 3.940 | 4.865 | 15.987 | 18.307 | 20.483 | 23.209 | 25.188 |
| **11** | 2.603 | 3.053 | 3.816 | 4.575 | 5.578 | 17.275 | 19.675 | 21.920 | 24.725 | 26.757 |
| **12** | 3.074 | 3.571 | 4.404 | 5.226 | 6.304 | 18.549 | 21.026 | 23.337 | 26.217 | 28.300 |
| **13** | 3.565 | 4.107 | 5.009 | 5.892 | 7.041 | 19.812 | 22.362 | 24.736 | 27.688 | 29.819 |
| **14** | 4.075 | 4.660 | 5.629 | 6.571 | 7.790 | 21.064 | 23.685 | 26.119 | 29.141 | 31.319 |
| **15** | 4.601 | 5.229 | 6.262 | 7.261 | 8.547 | 22.307 | 24.996 | 27.488 | 30.578 | 32.801 |
| **16** | 5.142 | 5.812 | 6.908 | 7.962 | 9.312 | 23.542 | 26.296 | 28.845 | 32.000 | 34.267 |
| **17** | 5.697 | 6.408 | 7.564 | 8.672 | 10.085 | 24.769 | 27.587 | 30.191 | 33.409 | 35.718 |
| **18** | 6.265 | 7.015 | 8.231 | 9.390 | 10.865 | 25.989 | 28.869 | 31.526 | 34.805 | 37.156 |
| **19** | 6.844 | 7.633 | 8.907 | 10.117 | 11.651 | 27.204 | 30.144 | 32.852 | 36.191 | 38.582 |
| **20** | 7.434 | 8.260 | 9.591 | 10.851 | 12.443 | 28.412 | 31.410 | 34.170 | 37.566 | 39.997 |
| **21** | 8.034 | 8.897 | 10.283 | 11.591 | 13.240 | 29.615 | 32.671 | 35.479 | 38.932 | 41.401 |
| **22** | 8.643 | 9.542 | 10.982 | 12.338 | 14.041 | 30.813 | 33.924 | 36.781 | 40.289 | 42.796 |
| **23** | 9.260 | 10.196 | 11.689 | 13.091 | 14.848 | 32.007 | 35.172 | 38.076 | 41.638 | 44.181 |
| **24** | 9.886 | 10.856 | 12.401 | 13.848 | 15.659 | 33.196 | 36.415 | 39.364 | 42.980 | 45.558 |
| **25** | 10.520 | 11.524 | 13.120 | 14.611 | 16.473 | 34.382 | 37.652 | 40.646 | 44.314 | 46.928 |
| **26** | 11.160 | 12.198 | 13.844 | 15.379 | 17.292 | 35.563 | 38.885 | 41.923 | 45.642 | 48.290 |
| **27** | 11.808 | 12.878 | 14.573 | 16.151 | 18.114 | 36.741 | 40.113 | 43.195 | 46.963 | 49.645 |
| **28** | 12.461 | 13.565 | 15.308 | 16.928 | 18.939 | 37.916 | 41.337 | 44.461 | 48.278 | 50.994 |
| **29** | 13.121 | 14.256 | 16.047 | 17.708 | 19.768 | 39.087 | 42.557 | 45.722 | 49.588 | 52.335 |
| **30** | 13.787 | 14.953 | 16.791 | 18.493 | 20.599 | 40.256 | 43.773 | 46.979 | 50.892 | 53.672 |
| **40** | 20.707 | 22.164 | 24.433 | 26.509 | 29.051 | 51.805 | 55.758 | 59.342 | 63.691 | 66.766 |
| **50** | 27.991 | 29.707 | 32.357 | 34.764 | 37.689 | 63.167 | 67.505 | 71.420 | 76.154 | 79.490 |
| **75** | 47.206 | 49.475 | 52.942 | 56.054 | 59.795 | 91.061 | 96.217 | 100.839 | 106.393 | 110.285 |
| **100** | 67.328 | 70.065 | 74.222 | 77.929 | 82.358 | 118.498 | 124.342 | 129.561 | 135.807 | 140.170 |

Table of the F-statistic for a probability level of 0.01 (Denominator to left, Numerator across top)

| d.f. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 20 | 30 | 40 | 50 | 75 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 161 | 199 | 5404 | 5624 | 5764 | 5859 | 5928 | 5981 | 6022 | 6056 | 6209 | 6260 | 6286 | 6302 | 6324 | 6334 |
| **2** | 18.51 | 19.00 | 99.16 | 99.25 | 99.30 | 99.33 | 99.36 | 99.38 | 99.39 | 99.40 | 99.45 | 99.47 | 99.48 | 99.48 | 99.48 | 99.49 |
| **3** | 10.13 | 9.55 | 29.46 | 28.71 | 28.24 | 27.91 | 27.67 | 27.49 | 27.34 | 27.23 | 26.69 | 26.50 | 26.41 | 26.35 | 26.28 | 26.24 |
| **4** | 7.71 | 6.94 | 16.69 | 15.98 | 15.52 | 15.21 | 14.98 | 14.80 | 14.66 | 14.55 | 14.02 | 13.84 | 13.75 | 13.69 | 13.61 | 13.58 |
| **5** | 6.61 | 5.79 | 12.06 | 11.39 | 10.97 | 10.67 | 10.46 | 10.29 | 10.16 | 10.05 | 9.55 | 9.38 | 9.29 | 9.24 | 9.17 | 9.13 |
| **6** | 5.99 | 5.14 | 9.78 | 9.15 | 8.75 | 8.47 | 8.26 | 8.10 | 7.98 | 7.87 | 7.40 | 7.23 | 7.14 | 7.09 | 7.02 | 6.99 |
| **7** | 5.59 | 4.74 | 8.45 | 7.85 | 7.46 | 7.19 | 6.99 | 6.84 | 6.72 | 6.62 | 6.16 | 5.99 | 5.91 | 5.86 | 5.79 | 5.75 |
| **8** | 5.32 | 4.46 | 7.59 | 7.01 | 6.63 | 6.37 | 6.18 | 6.03 | 5.91 | 5.81 | 5.36 | 5.20 | 5.12 | 5.07 | 5.00 | 4.96 |
| **9** | 5.12 | 4.26 | 6.99 | 6.42 | 6.06 | 5.80 | 5.61 | 5.47 | 5.35 | 5.26 | 4.81 | 4.65 | 4.57 | 4.52 | 4.45 | 4.41 |
| **10** | 4.96 | 4.10 | 6.55 | 5.99 | 5.64 | 5.39 | 5.20 | 5.06 | 4.94 | 4.85 | 4.41 | 4.25 | 4.17 | 4.12 | 4.05 | 4.01 |
| **11** | 4.84 | 3.98 | 6.22 | 5.67 | 5.32 | 5.07 | 4.89 | 4.74 | 4.63 | 4.54 | 4.10 | 3.94 | 3.86 | 3.81 | 3.74 | 3.71 |
| **12** | 4.75 | 3.89 | 5.95 | 5.41 | 5.06 | 4.82 | 4.64 | 4.50 | 4.39 | 4.30 | 3.86 | 3.70 | 3.62 | 3.57 | 3.50 | 3.47 |
| **13** | 4.67 | 3.81 | 5.74 | 5.21 | 4.86 | 4.62 | 4.44 | 4.30 | 4.19 | 4.10 | 3.66 | 3.51 | 3.43 | 3.38 | 3.31 | 3.27 |
| **14** | 4.60 | 3.74 | 5.56 | 5.04 | 4.69 | 4.46 | 4.28 | 4.14 | 4.03 | 3.94 | 3.51 | 3.35 | 3.27 | 3.22 | 3.15 | 3.11 |
| **15** | 4.54 | 3.68 | 5.42 | 4.89 | 4.56 | 4.32 | 4.14 | 4.00 | 3.89 | 3.80 | 3.37 | 3.21 | 3.13 | 3.08 | 3.01 | 2.98 |
| **16** | 4.49 | 3.63 | 5.29 | 4.77 | 4.44 | 4.20 | 4.03 | 3.89 | 3.78 | 3.69 | 3.26 | 3.10 | 3.02 | 2.97 | 2.90 | 2.86 |
| **17** | 4.45 | 3.59 | 5.19 | 4.67 | 4.34 | 4.10 | 3.93 | 3.79 | 3.68 | 3.59 | 3.16 | 3.00 | 2.92 | 2.87 | 2.80 | 2.76 |
| **18** | 4.41 | 3.55 | 5.09 | 4.58 | 4.25 | 4.01 | 3.84 | 3.71 | 3.60 | 3.51 | 3.08 | 2.92 | 2.84 | 2.78 | 2.71 | 2.68 |
| **19** | 4.38 | 3.52 | 5.01 | 4.50 | 4.17 | 3.94 | 3.77 | 3.63 | 3.52 | 3.43 | 3.00 | 2.84 | 2.76 | 2.71 | 2.64 | 2.60 |
| **20** | 4.35 | 3.49 | 4.94 | 4.43 | 4.10 | 3.87 | 3.70 | 3.56 | 3.46 | 3.37 | 2.94 | 2.78 | 2.69 | 2.64 | 2.57 | 2.54 |
| **21** | 4.32 | 3.47 | 4.87 | 4.37 | 4.04 | 3.81 | 3.64 | 3.51 | 3.40 | 3.31 | 2.88 | 2.72 | 2.64 | 2.58 | 2.51 | 2.48 |
| **22** | 4.30 | 3.44 | 4.82 | 4.31 | 3.99 | 3.76 | 3.59 | 3.45 | 3.35 | 3.26 | 2.83 | 2.67 | 2.58 | 2.53 | 2.46 | 2.42 |
| **23** | 4.28 | 3.42 | 4.76 | 4.26 | 3.94 | 3.71 | 3.54 | 3.41 | 3.30 | 3.21 | 2.78 | 2.62 | 2.54 | 2.48 | 2.41 | 2.37 |
| **24** | 4.26 | 3.40 | 4.72 | 4.22 | 3.90 | 3.67 | 3.50 | 3.36 | 3.26 | 3.17 | 2.74 | 2.58 | 2.49 | 2.44 | 2.37 | 2.33 |
| **25** | 4.24 | 3.39 | 4.68 | 4.18 | 3.85 | 3.63 | 3.46 | 3.32 | 3.22 | 3.13 | 2.70 | 2.54 | 2.45 | 2.40 | 2.33 | 2.29 |
| **26** | 4.23 | 3.37 | 4.64 | 4.14 | 3.82 | 3.59 | 3.42 | 3.29 | 3.18 | 3.09 | 2.66 | 2.50 | 2.42 | 2.36 | 2.29 | 2.25 |
| **27** | 4.21 | 3.35 | 4.60 | 4.11 | 3.78 | 3.56 | 3.39 | 3.26 | 3.15 | 3.06 | 2.63 | 2.47 | 2.38 | 2.33 | 2.26 | 2.22 |
| **28** | 4.20 | 3.34 | 4.57 | 4.07 | 3.75 | 3.53 | 3.36 | 3.23 | 3.12 | 3.03 | 2.60 | 2.44 | 2.35 | 2.30 | 2.23 | 2.19 |
| **29** | 4.18 | 3.33 | 4.54 | 4.04 | 3.73 | 3.50 | 3.33 | 3.20 | 3.09 | 3.00 | 2.57 | 2.41 | 2.33 | 2.27 | 2.20 | 2.16 |
| **30** | 4.17 | 3.32 | 4.51 | 4.02 | 3.70 | 3.47 | 3.30 | 3.17 | 3.07 | 2.98 | 2.55 | 2.39 | 2.30 | 2.25 | 2.17 | 2.13 |
| **40** | 4.08 | 3.23 | 4.31 | 3.83 | 3.51 | 3.29 | 3.12 | 2.99 | 2.89 | 2.80 | 2.37 | 2.20 | 2.11 | 2.06 | 1.98 | 1.94 |
| **50** | 4.03 | 3.18 | 4.20 | 3.72 | 3.41 | 3.19 | 3.02 | 2.89 | 2.78 | 2.70 | 2.27 | 2.10 | 2.01 | 1.95 | 1.87 | 1.82 |
| **75** | 3.97 | 3.12 | 4.05 | 3.58 | 3.27 | 3.05 | 2.89 | 2.76 | 2.65 | 2.57 | 2.13 | 1.96 | 1.87 | 1.81 | 1.72 | 1.67 |
| **100** | 3.94 | 3.09 | 3.98 | 3.51 | 3.21 | 2.99 | 2.82 | 2.69 | 2.59 | 2.50 | 2.07 | 1.89 | 1.80 | 1.74 | 1.65 | 1.60 |

Table of the F-statistic for a significance level of 0.05 (Denominator to left, Numerator across top)

| d.f. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 20 | 30 | 40 | 50 | 75 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 161.45 | 199.50 | 215.71 | 224.58 | 230.16 | 233.99 | 236.77 | 238.88 | 240.54 | 241.88 | 248.02 | 250.10 | 251.14 | 251.77 | 252.62 | 253.04 |
| **2** | 18.51 | 19.00 | 19.16 | 19.25 | 19.30 | 19.33 | 19.35 | 19.37 | 19.38 | 19.40 | 19.45 | 19.46 | 19.47 | 19.48 | 19.48 | 19.49 |
| **3** | 10.13 | 9.55 | 9.28 | 9.12 | 9.01 | 8.94 | 8.89 | 8.85 | 8.81 | 8.79 | 8.66 | 8.62 | 8.59 | 8.58 | 8.56 | 8.55 |
| **4** | 7.71 | 6.94 | 6.59 | 6.39 | 6.26 | 6.16 | 6.09 | 6.04 | 6.00 | 5.96 | 5.80 | 5.75 | 5.72 | 5.70 | 5.68 | 5.66 |
| **5** | 6.61 | 5.79 | 5.41 | 5.19 | 5.05 | 4.95 | 4.88 | 4.82 | 4.77 | 4.74 | 4.56 | 4.50 | 4.46 | 4.44 | 4.42 | 4.41 |
| **6** | 5.99 | 5.14 | 4.76 | 4.53 | 4.39 | 4.28 | 4.21 | 4.15 | 4.10 | 4.06 | 3.87 | 3.81 | 3.77 | 3.75 | 3.73 | 3.71 |
| **7** | 5.59 | 4.74 | 4.35 | 4.12 | 3.97 | 3.87 | 3.79 | 3.73 | 3.68 | 3.64 | 3.44 | 3.38 | 3.34 | 3.32 | 3.29 | 3.27 |
| **8** | 5.32 | 4.46 | 4.07 | 3.84 | 3.69 | 3.58 | 3.50 | 3.44 | 3.39 | 3.35 | 3.15 | 3.08 | 3.04 | 3.02 | 2.99 | 2.97 |
| **9** | 5.12 | 4.26 | 3.86 | 3.63 | 3.48 | 3.37 | 3.29 | 3.23 | 3.18 | 3.14 | 2.94 | 2.86 | 2.83 | 2.80 | 2.77 | 2.76 |
| **10** | 4.96 | 4.10 | 3.71 | 3.48 | 3.33 | 3.22 | 3.14 | 3.07 | 3.02 | 2.98 | 2.77 | 2.70 | 2.66 | 2.64 | 2.60 | 2.59 |
| **11** | 4.84 | 3.98 | 3.59 | 3.36 | 3.20 | 3.09 | 3.01 | 2.95 | 2.90 | 2.85 | 2.65 | 2.57 | 2.53 | 2.51 | 2.47 | 2.46 |
| **12** | 4.75 | 3.89 | 3.49 | 3.26 | 3.11 | 3.00 | 2.91 | 2.85 | 2.80 | 2.75 | 2.54 | 2.47 | 2.43 | 2.40 | 2.37 | 2.35 |
| **13** | 4.67 | 3.81 | 3.41 | 3.18 | 3.03 | 2.92 | 2.83 | 2.77 | 2.71 | 2.67 | 2.46 | 2.38 | 2.34 | 2.31 | 2.28 | 2.26 |
| **14** | 4.60 | 3.74 | 3.34 | 3.11 | 2.96 | 2.85 | 2.76 | 2.70 | 2.65 | 2.60 | 2.39 | 2.31 | 2.27 | 2.24 | 2.21 | 2.19 |
| **15** | 4.54 | 3.68 | 3.29 | 3.06 | 2.90 | 2.79 | 2.71 | 2.64 | 2.59 | 2.54 | 2.33 | 2.25 | 2.20 | 2.18 | 2.14 | 2.12 |
| **16** | 4.49 | 3.63 | 3.24 | 3.01 | 2.85 | 2.74 | 2.66 | 2.59 | 2.54 | 2.49 | 2.28 | 2.19 | 2.15 | 2.12 | 2.09 | 2.07 |
| **17** | 4.45 | 3.59 | 3.20 | 2.96 | 2.81 | 2.70 | 2.61 | 2.55 | 2.49 | 2.45 | 2.23 | 2.15 | 2.10 | 2.08 | 2.04 | 2.02 |
| **18** | 4.41 | 3.55 | 3.16 | 2.93 | 2.77 | 2.66 | 2.58 | 2.51 | 2.46 | 2.41 | 2.19 | 2.11 | 2.06 | 2.04 | 2.00 | 1.98 |
| **19** | 4.38 | 3.52 | 3.13 | 2.90 | 2.74 | 2.63 | 2.54 | 2.48 | 2.42 | 2.38 | 2.16 | 2.07 | 2.03 | 2.00 | 1.96 | 1.94 |
| **20** | 4.35 | 3.49 | 3.10 | 2.87 | 2.71 | 2.60 | 2.51 | 2.45 | 2.39 | 2.35 | 2.12 | 2.04 | 1.99 | 1.97 | 1.93 | 1.91 |
| **21** | 4.32 | 3.47 | 3.07 | 2.84 | 2.68 | 2.57 | 2.49 | 2.42 | 2.37 | 2.32 | 2.10 | 2.01 | 1.96 | 1.94 | 1.90 | 1.88 |
| **22** | 4.30 | 3.44 | 3.05 | 2.82 | 2.66 | 2.55 | 2.46 | 2.40 | 2.34 | 2.30 | 2.07 | 1.98 | 1.94 | 1.91 | 1.87 | 1.85 |
| **23** | 4.28 | 3.42 | 3.03 | 2.80 | 2.64 | 2.53 | 2.44 | 2.37 | 2.32 | 2.27 | 2.05 | 1.96 | 1.91 | 1.88 | 1.84 | 1.82 |
| **24** | 4.26 | 3.40 | 3.01 | 2.78 | 2.62 | 2.51 | 2.42 | 2.36 | 2.30 | 2.25 | 2.03 | 1.94 | 1.89 | 1.86 | 1.82 | 1.80 |
| **25** | 4.24 | 3.39 | 2.99 | 2.76 | 2.60 | 2.49 | 2.40 | 2.34 | 2.28 | 2.24 | 2.01 | 1.92 | 1.87 | 1.84 | 1.80 | 1.78 |
| **26** | 4.23 | 3.37 | 2.98 | 2.74 | 2.59 | 2.47 | 2.39 | 2.32 | 2.27 | 2.22 | 1.99 | 1.90 | 1.85 | 1.82 | 1.78 | 1.76 |
| **27** | 4.21 | 3.35 | 2.96 | 2.73 | 2.57 | 2.46 | 2.37 | 2.31 | 2.25 | 2.20 | 1.97 | 1.88 | 1.84 | 1.81 | 1.76 | 1.74 |
| **28** | 4.20 | 3.34 | 2.95 | 2.71 | 2.56 | 2.45 | 2.36 | 2.29 | 2.24 | 2.19 | 1.96 | 1.87 | 1.82 | 1.79 | 1.75 | 1.73 |
| **29** | 4.18 | 3.33 | 2.93 | 2.70 | 2.55 | 2.43 | 2.35 | 2.28 | 2.22 | 2.18 | 1.94 | 1.85 | 1.81 | 1.77 | 1.73 | 1.71 |
| **30** | 4.17 | 3.32 | 2.92 | 2.69 | 2.53 | 2.42 | 2.33 | 2.27 | 2.21 | 2.16 | 1.93 | 1.84 | 1.79 | 1.76 | 1.72 | 1.70 |
| **40** | 4.08 | 3.23 | 2.84 | 2.61 | 2.45 | 2.34 | 2.25 | 2.18 | 2.12 | 2.08 | 1.84 | 1.74 | 1.69 | 1.66 | 1.61 | 1.59 |
| **50** | 4.03 | 3.18 | 2.79 | 2.56 | 2.40 | 2.29 | 2.20 | 2.13 | 2.07 | 2.03 | 1.78 | 1.69 | 1.63 | 1.60 | 1.55 | 1.52 |
| **75** | 3.97 | 3.12 | 2.73 | 2.49 | 2.34 | 2.22 | 2.13 | 2.06 | 2.01 | 1.96 | 1.71 | 1.61 | 1.55 | 1.52 | 1.47 | 1.44 |
| **100** | 3.94 | 3.09 | 2.70 | 2.46 | 2.31 | 2.19 | 2.10 | 2.03 | 1.97 | 1.93 | 1.68 | 1.57 | 1.52 | 1.48 | 1.42 | 1.39 |

## 1.6 Hypothesis Testing:

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

1. State the significance level
2. State the null hypothesis $H_0$ and its alternative $H_1$
3. State the statistic used
4. State the critical region
5. Evaluate the statistic and state the conclusion

To be honest with yourself, you need to state what level of uncertainty is acceptable before you compute any statistics. People usually choose 95% or 99% certainty. In the first case you are accepting a one in twenty chance of accepting the hypothesis wrongly - a type II error. (Type I error – false positive - you reject the null hypothesis incorrectly. Type II error - false negative - null hypothesis is not rejected but its alternative H1 is actually true). If you compute the statistic and then state what significance level it passes (e.g. 80%), then you are a mush-headed scoundrel, your use of statistics means nothing, and you should be ashamed.

Proper construction of the null hypothesis and its alternative is critical to the meaning of statistical significance testing. Careful logic must be employed to ensure that the null hypothesis is reasonable and that its rejection leads uniquely to its alternative. 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. Usually the null hypothesis and its alternative are mutually exclusive. Examples follow.

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

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

$H_0$: The variance at a period of 5 days is less than or equal to the red-noise background spectrum
$H_1$: The variance at a period of 5 days exceeds the red-noise background spectrum

## 1.6b Errors in Hypothesis testing

Even though you have applied a test and the test gives a result, you can still be wrong, since you are making only a probabilistic statement. The following table illustrates the Type I: You reject the null hypothesis, but the true value is in the acceptance level, and Type II: You fail to reject the null hypothesis, but the true value is outside the acceptance level for the null hypothesis Ho. In the table below positive means that you reject Ho and find something interesting. Negative means you cannot reject Ho. Yes, it's confusing.

| | Ho is true ($e.g.$ $z < z_{crit}$) | Ho is false ($e.g.$ $z > z_{crit}$) |
|---|---|---|
| Fail to reject Null Hypothesis Ho | No Error<br>True negative | Type II Error<br>False negative |
| Reject Null Hypothesis Ho | Type I Error<br>False positive | No Error<br>True positive |

**First Example:**

In a sample of 10 winters the mean January temperature is 42˚F and the standard deviation is 5˚F. What are the 95% confidence limits on the true mean January temperature?

1. Desired confidence level is 95%.

2. The null hypothesis is that the true mean is between $42 \pm \Delta T$. The alternative is that it is outside this region.

3. We will use the $t$ statistic.

4. The critical region is $|t| < t_{.025}$, which for $n = N - 1 = 9$ is $|t| < 2.26$. Stated in terms of confidence limits on the mean we have:

$$\bar{x} - 2.26 \cdot \frac{s}{\sqrt{N-1}} < \mu < \bar{x} + 2.26 \cdot \frac{s}{\sqrt{N-1}}$$

5. Putting in the numbers we get $38.23 < \mu < 45.77$. We have 95% certainty that the true mean lies between these values. This is the answer we wanted. If we had a guess about what the true mean was, we could say whether the data would allow us to reject this null hypothesis at the significance level stated.

**Another Example: Dilbert's Ski Vacation**

Dilbert goes skiing 10 times at Snoqualmie Pass this season. The average temperature on the days he is there is 35˚F, whereas the climatological mean for the same period is 32˚F and the standard deviation is 5˚F. Has he been unlucky?

Well, let's suppose being unlucky means that the temps were warmer on the days he was there than average, and lets use (1) a significance level of 95%. The null hypothesis (2) is that the mean for his 10 days is no different than climatology, and its alternative is that his days were warmer, which would make him unlucky. The null hypothesis is that he is not unlucky. We will use (3) the *t*-statistic, for which the two-sided critical region (4) is $t > t_{0.025}$, which for 9 degrees of freedom is $t > t_{0.025} = 2.262$. Now we evaluate the statistic and state the conclusion (5),

$t = \frac{35-32}{5}\sqrt{9} = 1.80$ , which is less than the *t* required to reject the null hypothesis, so we cannot conclude that Dilbert was unlucky. His mean temperature is below the maximum that you would expect from a random sample of 10 at the 95% level. Notice how your conclusions about life depend on the level of certainty that you require. If Dilbert were less scientific, he might have concluded that he was unlucky just because the sample mean of his visits was greater than the climatological average. It is reassuring to Dilbert to realize that life is probabilistic, and that his experience is within the expected range. Of course if he had a good weather forecast, Dilbert could have planned his ski trips for colder weather and better snow, which would not make him lucky, but good.

## 1.7 *Combinatorics and the Binomial and Hypergeometric Distributions*

The number of ordered permutations of length k that can be drawn from a set of n distinct elements, repetition not allowed, is:

$$n(n-1)(n-2)\ldots(n-k+1) = \frac{n!}{(n-k)!} \tag{1.43}$$

If the order is not important, the number of combinations that can be drawn is:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{1.44}$$

The symbol $\binom{n}{k}$ gives the binomial coefficients. They are related to the following algebra problem, too.

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} \tag{1.45}$$

***Example:*** Eight people gather for dinner. If each person shakes hands with every other person only once, how many handshakes are required?

This is like an unordered draw of pairs from an urn containing 8 unique tokens. So the answer is:

$$\text{Handshakes} = \binom{8}{2} = \frac{8!}{2!\,(8-2)!} = 28 \textit{ handshakes are required}$$

You can also do this calculation by having one person shake hands with everyone(7 handshakes), then remove that person from the pool and pick another person (6 handshakes), and so on giving, 7+6+5+4+3+2+1 = 28 handshakes. If you are the host you can make this efficient by having your guests queue up outside. Then you shake each of their hands as they enter the room. The first person in line takes a place behind you in the reception line and shakes hands with everyone behind them in line, and so forth. The last person in line shakes hands with all the other folks now in the receiving line.

**Hypergeometric Distribution:**

If you have a bucket with *r* red balls and *w* white balls, so that *r+w=N*, the total number of balls in the bucket. If you draw *n* balls out of the bucket at random, then the probability of getting *k* red balls is:

$$P(k) == \frac{\binom{r}{k}\binom{w}{n-k}}{\binom{N}{n}} \quad ;\ \max(0,\text{n-w}) \le \text{k} \le \min(\text{n,r}) \tag{1.46}$$

**Binomial Distribution:**

Suppose you have a set of n trials in which the outcome is either "success" or "failure". The probability of success in one trial is p=P(success in one trial). If X is the total number of successes in n trials, then:

$$P(X=k)=\binom{n}{k} p^k\ (1-p)^{n-k}\ ,\ k=1,2,3,...n \qquad (1.47)$$

**Example:** What is the probability of getting more than 15 heads in 20 tosses of a fair coin? Answer:

$$\sum_{k=16}^{20}\binom{20}{k} 0.5^k\ (1-0.5)^{20-k}=0.006$$

The binomial distribution is helpful in assessing "field significance", the significance of multiple tests as when an array of variables are tested against the same hypothesis. An example would be correlating the sunspot index with a map of pressure at many points over the earth. How many individual "significant" events do you expect to get by chance in such cases. As an example, consider the plot below, which shows for N tries of a test at the p=0.05 significance level, what the binomial distribution (1.47) says about how many you should get by chance.

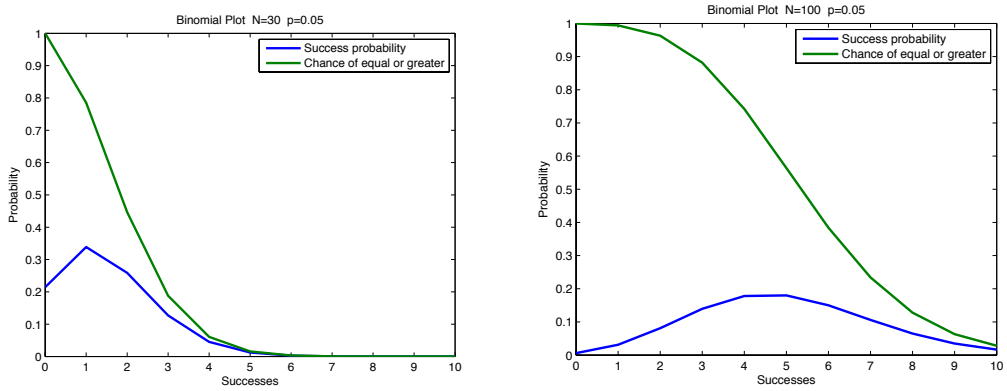



Figure Blue line shows probability of a given number of successes in N tries (N=30 on left and N=100 on right) where the probability of each event succeeding is p=0.05. Green line shows the probability of getting the number of successes or more.

Note that the probability of getting 5 successes or more in 30 tries is less than 0.05 and getting 10 successes or more in 100 tries is less than 0.05. That is 16.7% are successes for 30 tries and only 10% are successes for 100 tries at same probability level. For smaller samples, the fraction of total tries that can succeed by chance is greater. Even for 100 tries, 10% can succeed by chance, where the probability of each individual occurrence is p=5%. The most likely outcome is shown by the peak of the blue line and is what you expect, about 5% of the chances will succeed. But the chances of getting significantly more than that are quite good, and ten or fifteen percent of the field points could succeed by chance at the 5% level (See Wilks 2006, and Livezey and Chen 1983).

If you did the calculations above by hand you would find it tedious. This gets worse when the sample is larger. To assist in this it is useful to use the following Normal Approximation to the Binomial.

**Normal Approximation to the Binomial (DeMoivre-Laplace)**

From the central limit theorem, it can be shown that the distribution of sample means approaches the Normal Distribution, even if the population from which the means are derived is not normally distributed. This is also true for the Binomial distribution, for which values have a probability of being either zero or one, but nothing else. The distribution of sample means from a binomial population is nonetheless normally distributed about its mean value of 0.5.

Here is a Java Applet that allows you to play with the Normal Approximation http://www.ruf.rice.edu/~lane/stat_sim/normal_approx/ . Try calculating the probability of getting more than 40 hits in a sample of 60 with P=0.5. That's 40 or more heads out of 60 coin tosses. You can compare the exact and approximate probabilities. Try this with small samples. For example, try 4 or more heads out of 6 tosses.

**DeMoivre-Laplace Theorem.**

*X* is a binomial variable defined on *n* independent trials each having success probability *p*. Then for any numbers *a* and *b*,

$$\lim_{n\to\infty} P\left( a < \frac{X-np}{\sqrt{np(1-p)}} < b \right) = \frac{1}{\sqrt{2\pi np(1-p)}} \int_a^b e^{-x^2/2} dx \qquad (1.48)$$

This means that the statistic, $\frac{X-np}{\sqrt{np(1-p)}}$ , has the Normal distribution. We can use this fact to simplify the solution of binomial problems, as illustrated in the example below.

**Example of the Normal approximation to the Binomial:** An earthquake forecaster has forecast 200 earthquakes. How many times in 200 trials must she be successful so we can say with 95% certainty that she has nonzero skill?

The null hypothesis is that she has no skill and the confidence level is 0.05, or 95%. We then want,

$$P(s > s^* \mid H_0) = 0.025 = \sum_{s=s^*}^{200} \left( \frac{200}{s} \right) \left( \frac{1}{2} \right)^s \left( 1 - \frac{1}{2} \right)^{200-s}$$

Solving this equation for s>s*, the number of occurrences necessary to leave only a 0.025 probability to the right, is extremely tedious to do by hand, which nobody these days would do. However, we can use the Normal approximation to the Binomial to convert this to the problem,

$$P(s > s^* \mid H_0) = P\left( \frac{s - np}{\sqrt{np(1-p)}} > \frac{s^* - np}{\sqrt{np(1-p)}} \right) = P\left( Z > \frac{s^* - np}{\sqrt{np(1-p)}} \right)$$

Now P(Z > 1.96) = 0.025 (two-tailed 95%), so we want,

$$\frac{s - np}{\sqrt{np(1-p)}} > 1.96, \; or \; s > 114$$

Where we have inserted $n$=200 and $p$=0.5 to get the numerical value shown. So to pass a no-skill test on a sample of this size, the forecaster must be right 57% of the time. Of course, this level of skill, while significantly different from zero, may not be practically useful.

## 1.8 *Non-Parametric Statistical Tests*

The statistical tests applied above mostly assume that the samples come from populations for which the statistical distributions are known, or assumed, a priori. We very often assume that the statistics we are testing are Normally distributed, so we can use the shape of the Normal distribution in our tests. Tests have also been developed that do not require the assumption of a theoretical distribution. These are called 'non-parametric' or 'distribution-free' statistical tests. This approach can be easily illustrated with the Signs Test.

*1.8.1 Signs Test - also called the Wilcoxon Test*

Suppose we have paired data $(x_i, y_i)$. We want to know if the data have a shift in mean location from set $x_i$ to set $y_i$. We have a suspicion that the data are not Normally distributed and we don't want to assume that they are. Our null hypothesis is that the means of the two sets are identical. The alternative is that they are not. We will rather formulate the problem in terms of the median, $\tilde{\mu}$.

$$H_0 : \ \tilde{\mu}_1 = \tilde{\mu}_2 \qquad H_1 : \ \tilde{\mu}_1 \neq \tilde{\mu}_2$$

Let's reformulate this in terms of a probability that $y_i$ is greater than $x_i$.

$$H_0 : \ P(y_i > x_i) = 0.5 \quad , \ H_1 : \ P(y_i > x_i) \neq 0.5$$

Next replace each pair with a signed integer equal to one according to the following rule:

$$y_i > x_i \rightarrow +1$$
$$y_i < x_i \rightarrow -1$$

If the median values of the two sets are the same, then plus and minus signs should be equally probable. Since we've taken the magnitude out of the problem, we can assume that the + and - correspond to binomially distributed 'success' and 'failure'. The probability of getting a certain number of + and - signs can be calculated from the binomial distribution (1.47).

**Example: Cloud Seeding Experiment** Ten pairs of very similar developing cumulus clouds were identified. One from each pair was seeded, and the other was not. Then the precipitation falling from the clouds later was measured with a radar. The data in the following table resulted:

| Cloud Pair | Precip. (untreated) | Precip. (treated) | $y_i > x_i$ ? |
|---|---|---|---|
| 1 | 10 | 12 | + |
| 2 | 6 | 8 | + |
| 3 | 48 | 10 | - |
| 4 | 3 | 7 | + |
| 5 | 5 | 6 | + |
| 6 | 52 | 4 | - |
| 7 | 12 | 14 | + |
| 8 | 2 | 8 | + |
| 9 | 17 | 29 | + |
| 10 | 8 | 9 | + |

So we get 8+ and 2-. Is this statistically significant at the 95% level, so that we can say the median values of the two samples are different?? So we plug into the binomial distribution to see what the chances of getting 8 sucesses in 10 tries.

$$P(k \geq 8) = \sum_{k=8}^{10} \binom{10}{k} \left(\frac{1}{2}\right)^k \left(1-\frac{1}{2}\right)^{10-k} = 0.055$$
$$P(k \leq 2) = \sum_{k=0}^{2} \binom{10}{k} \left(\frac{1}{2}\right)^k \left(1-\frac{1}{2}\right)^{10-k} = 0.055$$

Since if things were random the chance of getting two or less successes is equally probable we have to add these two probabilities in a two-sided test and we find that the probability of the result we got was P=0.11, which fails a 95% confidence test. We expect to toss 8 out of ten heads or tails about 11% of the time. So our sample needs to be bigger to do much with the data set.

There are many more complicated distribution-free tests that you can look up in textbooks, like the 'Wilcoxon signed rank test' and the 'Wilcoxon-Mann-Whitney test' (e.g. Mendenhall, et al 1990).

### *1.8.2 Rank Sum Test*

Another common and classical non-parametric test is the Rank-Sum Test, or Wilcoxon-Mann-Whitney Test. Suppose we have two samples $S_1$ and $S_2$, of sizes $n_1$ and $n_2$ that we want to test for location. Our null hypothesis is that they come from the same population with the same distributions, and we want to see if we can reject this H0. Combine them into a single sample $N=n_1+n_2$ and rank them from smallest (R=1 to largest R=N). Next compute the rank sums of each sample, which is the sum of the ranks for each subsample $S_1$ and $S_2$, which will be $R_1$ and $R_2$.

$$R_1 + R_2 = 1 + 2 + 3 + .. + N = n(n+1)/2$$

$R_1/n_1$ and $R_2/n_2$ should be similar if H0 is true and they are from the same population. Consider our particular values of $R_1$ and $R_2$ as drawn from a large number of possible random combinations from the sample N. There would be $N!/(n_1!\, n_2!)$ possible such combinations. We could try to do the calculation combinatorically, but it is easier to use the U statistic introduced by Mann-Whitney.

$$\begin{aligned} U_1 &= R_1 - \frac{n_1}{2}(n_1+1) \\ U_2 &= R_2 - \frac{n_2}{2}(n_2+1) \end{aligned} \tag{1.48}$$

Where, then, $U_1 + U_2 = = \frac{n_1 n_2}{2}$.

The U-statistic is approximately Normally distributed with mean and standard deviation,

$$\mu_U = \frac{n_1 n_2}{2}$$
$$\sigma_U = \left[ \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} \right]^{1/2} \tag{1.49}$$

With this theoretical mean and standard deviation, the statistical significance of U can then be tested with the standard cumulative normal distribution tables for F(z).

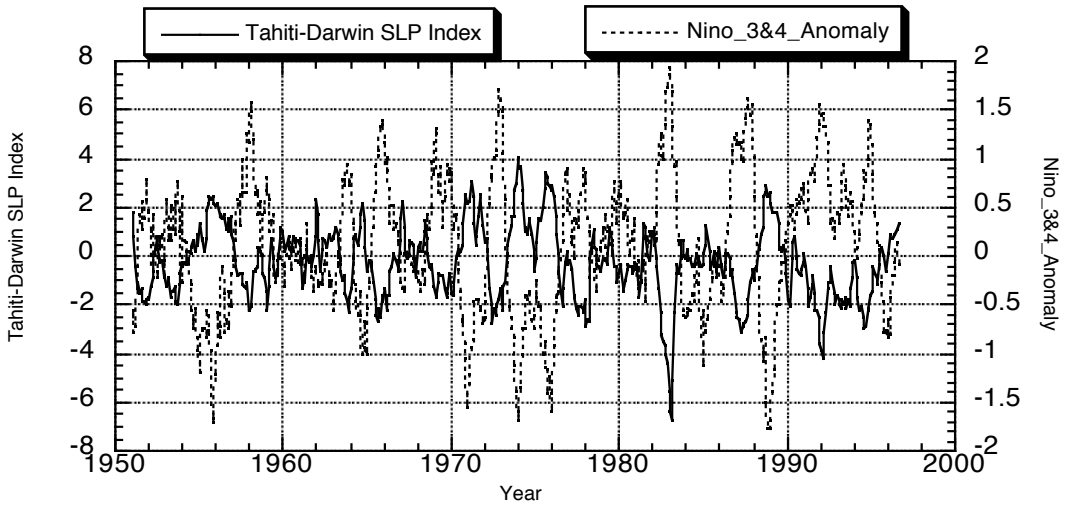
## 1.9 *A Priori, A Posteriori* and Degrees of Freedom

Applying statistical significance tests can be tricky, and it is very easy to get fooled. Two of the most difficult concepts are evaluating the true number of degrees of freedom in your data set, and knowing when you are entitled to use *a priori* statistical significance tests. These concepts are perhaps best illustrated with examples. Let's start with the number of degrees of freedom, then follow with an example of the *a posteriori* problem.

*Degrees of Freedom: The Autocorrelated Time Series Problem:*

The number of degrees of freedom is the number of independent measurements of the quantity or event of interest that is included in the sample. If we have a time or space series it is sometimes difficult to assess the number of independent realizations that we have, and the answer may depend on the time or space scale of the phenomenon of interest. Some quantitative techniques or rules of thumb have been developed for evaluation the number of degrees of freedom in spatial or temporal data sets. Discussion of these is best left until we have reviewed some background material in regression and time series analysis. If you can't wait, have a look at Leith(1973) and Bretherton et al.(1999). Below is an illustrative example.

Joe has the following data sets that are monthly anomalies of the Tahiti-Darwin sea level pressure difference and anomalies of SST over two regions in the equatorial eastern Pacific. The data run from 1951 to 1996, so he has 46 years of monthly data, or 46x12 = 552 data points. When he tests for the statistical significance of the correlation between these two time series, he uses 552 as the number of independent data. Is he right?




He has overestimated the number of degrees of freedom in his time series, because not all of the points are independent. The inherent time scale of the largest variations in these time series seems to be much longer than one month. If you did an auto-regression analysis of these time series you would find that you can predict each month's value pretty well by using the two values adjacent to it, the month before and the month after. A method of estimating the true number of degrees of freedom for an autocorrelated time series is discussed in Section 6.15 of these notes.

*How Many Chances: The a priori Problem:*

Next Joe decides to test which day in December has the most rain in Seattle. December is one of the rainiest months, but Joe wants to know if there is any day in the month with more rain than the others. To test this he calculates the mean and standard deviation for each day in December from a 120 year record. December precipitation is uncorrelated from day to day, pretty much, so he actually has 120 independent data points for each day in December. The standard deviations for each day are pretty similar, so he uses the grand mean standard deviation for his statistical significance tests. He tests for the difference between the mean for each day and the grand mean for all the days to see if any day stands out. He finds that the mean for December 10 exceeds the grand mean of all the days in the month sufficiently to pass a 99% confidence limit. He then hastily begins writing a paper and speculating on the reasons why Dec 10 should be the rainiest day of December, since it is more rainy than the other days at the 99% level. Is he right?

Of course Joe has missed something. He gave the daily mean precipitation 31 chances to exceed the 99% probability, since each of the 31 days of December is an independent sample drawn, presumably, from a very similar population. To estimate the chance of this you take the probability of one event exceeding the criterion, 99% and raise it to the power that is the number of independent chances you have given the events to exceed this probability. Here we are assuming each event is independent and using (1.10). If we go back to section 1.2 and take a look at (1.10), then we can calculate that

the probability that none of our 31 independent estimates will pass the 99% significance level is $(0.99)^{31} = 0.73$, so our 99% significance is really 73% significance, which is not a number on which most people would bet their life savings or their reputations. Of course, if a significance level of 95% had been chosen, . . . well I leave the calculation to you (20% confidence level, the odds are 5-1 that one day will exceed 95% confidence.).

In order to score with this analysis, Joe would have to have an *a priori* reason for expecting the 10th day of the month to be special. Usually this would require some reasonable theory for why this day would be special. In this case he has no *a priori* reason to think Dec. 10 is special, so he must use *a posteriori* statistical analysis, and his result is not good enough to reject the null hypothesis that all the days are drawn from the same population with the same mean and standard deviation. It is just too likely that one out of his 31 days would pass the 99% significance level by chance.

*Modify the Hypothesis:*

This time Joe decides that the solar cycle must have some control over weather. So he starts looking for correlations between weather variables and solar cycle variables. He tries three solar variables; sunspot number, solar diameter, and Lyman alpha flux. He correlates them all with 10 weather variables; global mean surface temperature, global mean precipitation, zonal mean wind speed, blocking frequency, . . . , tree ring width in Wyoming, vorticity area index. Finally he finds a correlation between solar diameter and vorticity area index (the area of the Northern Hemisphere with vorticity greater than 3 times the standard deviation above the average value) that is significant at the 95% level. He again begins to prepare a paper. Of course, we now know to comment that he tried all sorts of things before he tried the pair of things that pass 95%, and that he is not entitled to use *a priori* statistical tests. Of course, Joe may neglect to include in his paper the fact that he tried all this other stuff, and in fact maybe he didn't. Maybe someone else did, and he came up with the idea of vorticity area index against solar diameter after reading their papers. If people believe a relationship should exist between two classes of things and keep trying different alternatives for showing that relationship, their chances of eventually finding something that passes *a priori* tests is pretty good.

*Elaborate the Hypothesis*

Eventually, more data are gathered and Joe's correlation between solar diameter and vorticity area index falls apart. Undaunted, he decides that something about the weather has changed, perhaps global warming, and that this is significant. His paper to Nature on this topic is rejected. Then he goes back to the drawing board. He decides to divide the sample into years when the Quasi-Biennial Oscillation of the stratosphere is easterly and years when the QBO is westerly. To his satisfaction, the correlation returns to the 95% level for the westerly years and appears not at all in the easterly year sample. In this case he is committing the egregious sin of elaborating the hypothesis, or dividing the sample until he gets the result he wants, unless he has a good reason for expecting only the westerly years to produce the result. If he has no physical theory behind his

dividing and elaborating, then he is not being scientific and is likely to produce a lot of rubbish.

*Final Words on Data Exploration*

If you follow all the rules against applying *a priori* statistical tests to *a posteriori* conclusions, don't modify hypotheses, and don't elaborate hypotheses – you may never discover anything new through exploratory data analysis. Most of the interesting things in meteorology and oceanography have been observed first and then explained with theory later. Often the data used in the initial discovery are inadequate in quality or volume. Some combination of art and science is needed to find new things that are true using statistics as a tool. One needs some combination of firm grounding in probability and statistics, plus physical intuition, plus common sense, to be successful.

*Frequentist versus Bayesian Views of Statistics:*

Here we discuss the frequentist view of statistics, which is to approximate probabilities from observations using a large sample. This works if you have a large sample and the question you are asking is well defined. Bayesian statistics are another method that allows for taking into account prior information and working with small samples.

*Weak and Strong Laws of Large Numbers*

Suppose P(E) is the probability of some event E. $P(E^T) = 1 - P(E)$ is the probability of the complementary event $E^T$.

$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean based on size *n* for which the true mean is $\mu$.

Weak

$$\lim_{n \to \infty} \; P(\left|\overline{x}_n - \mu\right| < \varepsilon) = 1$$

Strong

$$P\left( \lim_{n \to \infty} \; \overline{x}_n = \mu \right) = 1$$

This indicates that we can get arbitrarily close to the true mean with a large enough sample.

For probability:

Suppose we have a random number $x_i$ that is either zero or one. $P(x_i = 1) = p$.

$$y_n = \sum_{i=1}^{n} x_i \quad \text{and then} \quad P\left(\left|\frac{y_n}{n} - p\right| < \varepsilon\right) \to 1 \qquad for\ \varepsilon > 0$$

That is, if you have a large enough sample, you can measure the true probability to an arbitrary precision. This is the basis of the frequentist view of probability.

*Uncertainty: What are statistics good for?*

Statistics test whether something could have occurred by chance, subject to some assumptions and prior assumptions. Statistics test only one kind of uncertainty, of which we can define three.

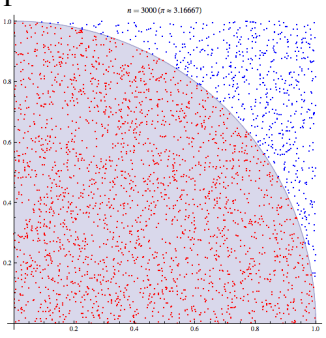
Aleatory Uncertainty: This is random uncertainty that we can measure with statistics.

Epistemic Uncertainty: This is uncertainty associated with lack of knowledge about things we could in principle know about. We know the physics, but are uncertain about the parameters.

Structural Uncertainty: This is uncertainty arising from things we don't know about. Unknown unknowns. We thought the world was flat, but it is really spherical.

## 2.0 Monte Carlo Methods

In the age of computers, sometimes it is easier to let the computer do the work, do many intelligently designed calculations, and then infer a fact or statistical conclusion from the aggregate of these calculations. The name Monte Carlo comes from the famous casino, not from the inventor of the method. It is a term that has no precise definition and covers a wide variety of techniques, which share in common the idea expressed in the first sentence. One famous example is the calculation of pi, the ratio of the circumference of a circle to its diameter. It can be calculated by inscribing a circle within a square, then drop pebbles randomly on the square. Count the ratio of the pebbles in the square to those that fall within the area. If the pebbles are dropped randomly, then this ratio should be the ratio of the areas of the circle to the square, which is pi/4. If you do this many times you can get an arbitrarily good approximation to pi.



Another example of a Monte Carlo method might be if a compositing procedure in which a division of the sample by some criterion leads to a particular quantitative result. For example, you take a selection of 10 items out of a sample of 25 using some criterion and get a measure that is different from some expectation. Take random samples of 10 from those 25 many, many times, and see how likely it is to get that measure or greater by chance. You can do this by either replacing each realization drawn, so that a given realization can appear more than once in a sample, or by removing the realization from the set of 25 when it is drawn so that it can appear only once in a subsample of 10. The

advantage of this method is that you don't have to choose a model PDF and you can evaluate the number of successes in exceeding the criteria using the binomial distribution.

Monte Carlo methods, defined broadly, are also being used in ensemble weather and climate predictions, where multiple numerical forecasts are done using initial conditions and/or forcings that vary within the range that they are known. The results of many such calculations can be used to assess the likelihood that the prediction will be reliable and often also provide a more precise definition of the most-likely reality. For example, averages of forecasts are often more accurate than any individual ensemble member.

Monte Carlo methods can also be used in strategy for games like Go or Chess. A program is given a little intelligence about chess playing. Individual moves are chosen randomly from a set of possible and plausible moves. The game is played many times to see which move produces the most likelihood of a good outcome. The move is made, the response is chosen and the simulation can be done again.

**References:**

Barnett, T. P., 1995: Monte-Carlo Climate Forecasting. *J. Climate*, **8,** 1005-1022.

Barnston, A. G., and R. E. Livezey, 1987: Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. *Mon. Wea. Rev.*, **115,** 1083-1126.

Elsner, J. B., X. F. Niu, and T. H. Jagger, 2004: Detecting shifts in hurricane rates using a Markov chain Monte Carlo approach. *J. Climate*, **17,** 2652-2666.

Jones, P. D., M. New, D. E. Parker, S. Martin, and I. G. Rigor, 1999: Surface air temperature and its changes over the past 150 years. *Rev. Geophysics*, **37,** 173-199.

Livezey, R. E., and W. Y. Chen, 1983: Statistical field significance and its determination by Monte-Carlo techniques. *Mon, Wea, Rev,*, **111,** 46-59.

Mears, C. A., F. J. Wentz, P. Thorne, and D. Bernie, 2011: Assessing uncertainty in estimates of atmospheric temperature changes from MSU and AMSU using a Monte-Carlo estimation technique. *J. Geophys. Rese.-Atmos.*, **116,** DOi:10.1029/2010jd014954.

Tomassini, L., P. Reichert, R. Knutti, T. F. Stocker, and M. E. Borsuk, 2007: Robust Bayesian uncertainty analysis of climate system properties using Markov chain Monte Carlo methods. *J. Climate*, **20,** 1239-1254.

Wilks, D. S., 2006: On "field significance" and the false discovery rate. *J. Appl. Meteor. Climat.*, **45,** 1181-1189.

Zhang, X. B., F. W. Zwiers, and G. L. Li, 2004: Monte Carlo experiments on the detection of trends in extreme values. *J. Climate*, **17,** 1945-1952.

**Exercises:**

1.1 What are the 95% confidence limits on the variance in the first example above?

1.2 What are the 95% confidence limits on the mean if you (wrongly) use the $z$ statistic in the first example?

1.3 In a composite of 10 volcanoes the average monthly mean temperature anomaly three months after the eruption is 0.5˚C colder than the monthly mean temperature anomaly one month before the eruption. The standard deviation of the monthly anomalies is 2.0˚C. Is this result significant?

1.4 The average snowfall at Stevens Pass is 415 inches. This has a standard deviation of 100 inches and is based on 30 yearly observations. What are the 99% confidence limits on the true mean?

1.5 Annual mean precipitation in Seattle averages 894 mm, with a standard deviation of 182 mm over a sample of 111 years. The 20 years from 1920-1939 had an average of 785mm and a standard deviation of 173mm. Were the 20's and 30's different from the overall record?

1.6 Of swordfish taken by the *Hannah Boden*, 0.5% have cancerous lesions. If the fish has cancerous lesions 90% also have high levels of PCP. If the fish do not have lesions, 30% nonetheless have high PCP. Use Bayes theorem to estimate how many of the fish with high PCP have lesions. Also, what is the probability that a swordfish has high PCP?

1.7 A student takes a 20-question multiple-choice exam where every question has 5 choices. Some of the answers she knows from study, some she guesses. If the conditional probability that she knows the answer, given that she answered it correctly, is 0.93, for how many of the questions did she know the answer?

1.8 An earthquake forecaster forecasts 20 earthquakes. How many of these forecasts must be successes in order to pass a 95% confidence that the forecaster has nonzero skill? What percentage is this? Compare to the case of 200 forecasts given above.

1.9 Historically, major midwinter warmings of the Northern Hemisphere stratosphere occur every other year (*i.e.* the probability of getting one in any year is p=0.5). During the decade of the 1990's, seven successive years without major warmings occurred. What is the probability of this event occurring by chance? What if it was not seven consecutive non-warming years, but merely 7 out of 10 in any order? What is the probability of getting 7 or more warmingless winters out of 10?

1.10 You are a contestant on *The Price is Right*. Monte Hall shows you three doors. Behind one door is a Tesla Roadster. Behind the other two doors are goats. After you choose your door, but before it is opened, Monte opens another door to reveal a goat. He offers you the opportunity to switch to the other unopened door. Should you switch? What is the probability that your first choice is the car? What is the probability that the alternative offered is the car? You can assume that Monte is

unbiased and does this goat reveal on every occasion. Explain your answer using statistics covered in this chapter.

1.11 You correlate a seasonal mean el Niño index against the seasonal anomalies of precipitation at 24 locations in the western USA. Eight of the twenty four locations show a relationship with el Niño that is significant at the 95% level. Assuming that each of the 24 locations is independent of the others, what is the probability that this result occurred by chance? Later you discover that the 24 stations are not independent. Using an objective method, you group the 24 stations into three groupings that are independent of each other. You find that one of these three groupings is significantly related to el Niño. How likely is it that this second result occurred by chance?

**Bibliography:**

Bretherton, C. S., M. Widmann, V. P. Dymnikov, J. M. Wallace and I. Bladé, 1999: The effective number of spatial degrees of freedom of a time-varying field. *Journal-of-Climate*, **12**, 1990-2009.

Howson, C. and P. Urbach, 2006: *Scientific Reasoning: The Bayesian Approach 3rd Ed.* Open Court, 470 pp.

Huff, Darrell, 1954: *How to Lie with Statistics*, Norton, 142.

Jaynes, E. T., 2003: *Probability Theory: The Logic of Science*. Cambridge University Press, 758 pp.

Knight, K., 2000: *Mathematical statistics. Texts in statistical science*, Chapman and Hall/CRC, 481 pp.

Larson, R. L. and M. L. Marx, 1986: *An Introduction to Mathematical Statistics and its Applications*. 2nd, Prentice-Hall, Englewood Cliffs, N.J., 630.

Leith, C. E., 1973: The standard error of time-averaged estimates of climatic means. *J. Appl. Meteorol.*, **12**, 1066-1069.

Mendenhall, W., D.D. Wackerly, and R.L. Sheaffer, 1990: Mathematical Statistics with Applications, PWS-Kent, Boston, p 688.

Panofsky, H. A. and G. W. Brier, 1968: *Some Applications of Statistics to Meteorology*. Pennsylvania State University, University Park, 224.

Wilks, D. S., 1995 and 2005: *Statistical Methods in the Atmospheric Sciences*. Academic Press, San Diego, 467.

Zwiers, F. W. and H. von Storch, 1995: Taking serial correlation into account in tests of the mean. *J. Climate*, **8**, 336-351.