Some Statistical Notions

a. Definitions

- **random variable** - a quantity which takes on random values from a distribution. **Example**: the year at which a particular day hits its maximum temperature value.
- **Sample Space** - the set of all outcomes of a random variable.
- **population** - The set of all possible outcomes or values of a random variable. **Example**: all possible rolls of a die (that result in a face showing!). That's an infinite space, by the way....
- **sample** - a subset of the population units. "Theory deals with populations, and practice deals with samples." **Example**: we roll the die thirty times.
- **probability density function (pdf)** - the distribution of probabilitiy for all possible values of a random variable (often denoted ρ) – that is, for the entire population. For a discrete variable, the probability of each potential value is given (and they sum to 1):

  Example: for a fair die, with a discrete distribution, $\rho(k) = 1/6$ (an example of a discrete uniform distribution).

  For a continuous variables, we talk about the probability of **intervals**, and require that $\rho(x) \geq 0$, and that

  $$\int_R \rho(x)dx = 1$$

  where R is the region over which ρ is defined (i.e., the random variable can take only values from the set R).

  Examples: Some important distributions:

  i. **uniform distribution**

  $$\rho(x) = \begin{cases} \frac{1}{\beta-\alpha} & \text{for } \alpha \leq x \leq \beta \\ 0 & \text{elsewhere} \end{cases}$$

  ii. **normal distribution**

  $$\rho(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

  where $\sigma > 0$ is the standard deviation, and μ is the mean. The dreaded bell-shaped curve.

  iii. **binomial distribution**

  $$(p+q)^n = p^n + np^{n-1}q + \ldots + \left(\begin{array}{c} n \\ k \end{array}\right) p^k q^{n-k} + \ldots + npq^{n-1} + q^n = 1$$

  where

  $$\left(\begin{array}{c} n \\ k \end{array}\right) = \frac{n!}{(n-k)!k!}$$

  since $q = 1 - p$. Thus we have the discrete binomial distribution

  $$\rho(k) = \left(\begin{array}{c} n \\ k \end{array}\right) p^k q^{n-k}$$

  for $k = 0, 1, \ldots, n$.

- **cumulative distribution function (cdf)** - $F(x) = P[X \leq x]$. We read that as "F of x is the probability that the random variable takes a value X, where $X \leq x$." **Example**: for a fair die, the cdf is a stairstep function:

$$cdf(x) = \begin{cases} 0 & k < 1 \\ \frac{floor[k]}{6} & 1 \leq k \leq 6 \\ 1 & k > 6 \end{cases}$$

  Example: for a normal

$$cdf(x) = \frac{1}{2}\left[1 + erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$$

- **expected value of** X - aka the mean of X

$$E(X) = \sum_k k\rho(k)$$

  for a discrete variable, and

$$E(X) = \int_{-\infty}^{\infty} x\rho(x)dx$$

  for a continuous variables. We often use $\mu = E(X)$.

  Example: So for a fair die, $\rho(k) = 1/6$, for $k \in \{1, 2, 3, 4, 5, 6\}$. The mean roll is thus

$$E(X) = \sum_{k=1}^{6} k\frac{1}{6} = 3.5$$

- **expected value of** $g(X)$ - where g is some function of the random variable X

$$E(g(X)) = \sum_k g(k)\rho(k)$$

  for a discrete variable, and

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)\rho(x)dx$$

  for a continuous variables.

  Example: For a fair die: $\rho(k) = 1/6$, for $k \in \{1, 2, 3, 4, 5, 6\}$. We might ask "What's the mean value of the square of X?" (i.e. $g(X) = X^2$). That would be

$$E(X) = \sum_{k=1}^{6} k^2\frac{1}{6} = (1 + 4 + 9 + 16 + 25 + 36)/6 = 15.1\overline{6}$$

- **variance of X** -

$$V(X) = E((X-\mu)^2) = E(X^2) - (E(X))^2$$

  Example: For the fair die, $V(X) = E((X-\mu)^2) = E(X^2) - (E(X))^2 = 15.1\overline{6} - 3.5^2 = 2.91\overline{6}$.

b. Theorems

- $$P[a \leq X \leq b] = \sum_{k=a}^{b} \rho(k)$$

  for a discrete variable, and
  $$P[a \leq X \leq b] = \int_a^b \rho(x) dx$$
  for a continuous random variable, where ρ is the pdf.

- Important calculus link: **The pdf is the derivative of the cdf.**

c. Summary

Upshot for us:

i. **Is climate change happening?**

   If not, then all climates would be the same. Every thirty year period would be the same, in terms of the number of extremes – of each type – that it contributed. The same would be true for every decade.

   So we would expect to see a **uniform** distribution of extreme-years-by-decade, or extreme-years-by-climate.

   We have rejected that in the Fletcher data.

ii. **Given that climate change is happening, how is it happening?**

   Our preliminary notion was that the temperature distribution across time is staying the same in terms of spread, but is rising in mean.

   For example, assume a normal distribution of temperatures, with sinusoidally varying mean
   $$\mu = 50.2 + 25 \sin 2\pi(t - t_0),$$
   and with $\sigma = 10$. "Wood County has an average of just above 50^o...." (50.2 he asserts on p. 64). As for deviations from the mean, he reports that "Temperature ranges remain around 50 degrees in the extreme with the fall months at or above that value, with May the only other month at that value. May has a wide spread in possible temperatures, from around freezing to the high '90s. The four warm months of June through September had only a five-degree variance, from 45 degrees to 50 degrees, while the reamining months showed only six degrees difference."(p. 68) So a $\sigma = 10$ would give a $\pm 3\sigma$ of 60. That seems fairly reasonable.

   Fletcher is arguing that σ should also vary seasonally. Perhaps we might consider something like
   $$\sigma(t) = 10 + \sin 2\pi(t - t_1)$$

Fletcher mentions that January, 1919 was a cold one, and had a mean temperature of 18.1 degrees. This jives with a value of $\mu = 20$ ($\mu - 3 * \sigma$), per our formula).

Now climate change means that this these μ and/or σ functions should change over time. Our first hypothesis was that we might replace μ with something like

$$\mu = 50.2 + ct + 25 \sin 2\pi(t - t_0),$$

where ct represents a constant increase in mean temperature over time.

We might try to construct (in this fashion) a set of distributions of temperatures so that if we simulate with these, over the course of 100 years or so, we generate distributions of MaxMax, MinMin, etc. which looks like the ones we actually find in Fletcher's data.