I need to postpone examples of these conditional expectation calculations until we've covered more standard distributions.

\[ Z, Y \text{ rv's such that } f_{Y|Z}(y|x) \text{ exists} \rightarrow \text{it makes sense to speak not only of } E(Z|X), \text{ the mean of } f_{Y|Z}(y|x), \text{ but also of the variance of that dist.}\]

\[ \text{Def } V(Z|X) = \mathbb{E}\left[\left(Y - E(Z|X)\right)^2 \mid X\right] \]

is called the conditional variance of \( Y \) given \( X = x \), and the rv \( V(Z|X) \) is just \( g(X) \), the conditional variance of \( Y \) given \( X \).
The payoff of this theorem relates to the constant $Z_0$ from $f$.

For all the losses, we can find that all of the following expressions exist. Let $E$.

Part (2) of the theorem states that the conditional expectation of $Y$ given $X$ is:

$$E(Y | X) = E(Y) + E(X) - E(X | Y)$$

For MSE $E((E - Z)^2)$ to predict $Y$ from $Z$, the condition $$(E - Z)^2$$

$$E(Y | X)^2 - E(X | Y)^2$$

Note that $E(Y)$ is $Z$. $E(X)^2$
Imagine a 2-port game.

**Stage 1**

Predict \( \hat{Y} \) without knowing \( \hat{X} \). Well, if you bury into MSE as your measure of "goodness" of a prediction, we know that you should predict \( \hat{Y} = \mu_y = E(Y) \).

And your resulting MSE will be

\[
E[(Y - \hat{Y})^2] = V(Y) = \sigma^2_Y
\]

**Stage 2**

Observe \( \hat{X} \).

Now predict \( \hat{Y} \).

Let's say \( \hat{X} = x^* \). Then we know the MSE-optimal prediction is

\[
\hat{Y} = \hat{E}(Y|X=x^*)
\]
and your resulting MSE will be

\[ E \left\{ [\epsilon - E(\theta | x^*)]^2 \right\} = V(\theta | x^*) \]

\[ \text{(**)*)} \]

From the vantage point of someone thinking about stage 2 before it happens, it is not yet known, so the expected value of \((**)*)\), namely \( E_\theta \left[ V(\theta | \Theta) \right] \), is the best you can do to guess at how good the stage 2 prediction will be. The second part of the double expectation theorem says

\[ V(\theta) = E_\theta \left[ V(\theta | \Theta) \right] + \frac{1}{n} \left( E(\hat{\theta} | \Theta) \right) \]

\[ \text{MSE of } \hat{\theta}_{\text{no \Theta}} \]

\[ \text{"E(MSE)" of } \hat{\theta}_{\text{no \Theta}} = E(\hat{\theta} | \Theta) \]
utility

calculate

Another complete solution is often to predict accuracy (or at least try the same) when you use \( \hat{f}(x) \) to get better (or at least stop the

Now you always expect your predictive

\[ E(\hat{f}(x)) \leq \text{RSE} \sqrt{\frac{2}{n}} \]

The

\[ V(\hat{f}(x)) + \frac{\text{RSE}^2}{n} \geq \frac{\text{RSE}^2}{n} \]

where the consequences are uncertain.
There is a theory of optimal action under uncertainty, it's called Bayesian decision theory - a concept called utility is central to this theory. The theory takes its simplest form when comparing gambles.

Example: \( X \) has discrete PMF \( f_X(x) = \begin{cases} \frac{1}{2} & x = -350 \\ \frac{1}{2} & x = +8500 \\ 0 & \text{else} \end{cases} \)

Suppose \( X \) is your net gain from gamble \( A \), and \( Y \) is your net gain from gamble \( B \).

So is \( A \) automatically better? 

\( E(X) = 775 \), \( E(Y) = -50 \). Then \( B \)?
Note that with \( B \) you're guaranteed to win at least 840, while \( A \) has no such guarantee; is \( A \) still automatically better for you than \( B \)? A risk-averse person would grab \( B \) quickly; a risk-seeking person would pick \( A \).

Evidently something more than just computing \( E(\xi) \), \( E(\eta) \) is going on.

**Def.** Your utility function \( U(x) \) of utility function is that function which assigns to each possible net gain \(-\infty < x < \infty \) a real \# \( U(x) \) representing the value to you of gaining \( x \).
if \( x \) is money, why not just use \( x \)?

\( u(x) = x \).

\( A: \) lovely, subtle answer first (utility is money)

Supplied by Daniel Bernoulli (1700 - 1782), a Swiss mathematician related to Jacob Bernoulli (1654 - 1705), for whom the Bernoulli distribution was named.

\( Daniel B: \) If your entire net worth is (say) $10, then the value to you of a new $1 is much greater than if your entire net worth is (say) $1,000,000; thus the utility of money is sublinear (meaning that it doesn't grow with \( x \) as fast as \( f(x) = x \) does).

\( Daniel B \) proposed one particular sublinear function for utility,
namely \( U(x) = 1 + \log(x) \) \( u(x) \) (for \( x > 0 \))

(Although Daniel B also invented the word utility)

(Although the idea goes back at least to Aristotle (384-322 BCE))

**Definition**

**Principle of Expected Utility Maximization**

You are said to choose between gambles by maximizing expected utility if, with \( U(x) \) your utility function, you prefer gamble \( X \) to gamble \( Z \) if \( E[U(X)] > E[U(Z)] \) and (\( \Box \)) you're indifferent between \( X \) and \( Z \) if \( E[U(X)] = E[U(Z)] \)
MEU first explored in depth by British philosopher economist Frank Ramsey (1903 - 1930), who died at 26 of liver failure (hepatitis).

Theorem (von Neumann - Morgenstern (1947): Under 4 reasonable axioms, MEU is the best you can do.

Suppose you bought a single $2 ticket in the PowerBall lottery examined in Take-Home Test 2: Suppose the drawing on 30 Jul 2016 for which the grand prize was $487 million. Let $X$ be the amount you will win (think: $8 before the drawing).
<table>
<thead>
<tr>
<th>Match</th>
<th>$x$</th>
<th>$P(x=x)$</th>
<th>$x \cdot P(x=x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5W, 1R</td>
<td>$487,000,000$</td>
<td>$\frac{1}{292,201,338}$</td>
<td>$1.667$</td>
</tr>
<tr>
<td>5W, 0R</td>
<td>$1,000$</td>
<td>$\frac{1}{11,688,053.52}$</td>
<td>$0.086$</td>
</tr>
<tr>
<td>4W, 1R</td>
<td>$850,000$</td>
<td>$\frac{1}{933,129.18}$</td>
<td>$0.055$</td>
</tr>
<tr>
<td>4W, 0R</td>
<td>$100$</td>
<td>$\frac{1}{1,352,351.77}$</td>
<td>$0.007$</td>
</tr>
<tr>
<td>3W, 1R</td>
<td>$100$</td>
<td>$\frac{1}{114,494.11}$</td>
<td>$0.009$</td>
</tr>
<tr>
<td>3W, 0R</td>
<td>$7$</td>
<td>$\frac{1}{59,760.00}$</td>
<td>$0.001$</td>
</tr>
<tr>
<td>2W, 1R</td>
<td>$7$</td>
<td>$\frac{1}{721.33}$</td>
<td>$0.001$</td>
</tr>
<tr>
<td>1W, 1R</td>
<td>$84$</td>
<td>$\frac{1}{91.98}$</td>
<td>$0.043$</td>
</tr>
<tr>
<td>0W, 1R</td>
<td>$4$</td>
<td>$\frac{1}{38.32}$</td>
<td>$0.104$</td>
</tr>
</tbody>
</table>

$\text{E}(x) = \sum \text{x} \cdot P(\text{x} = \text{x}) = \$1.99$

It has 9 possible values $x$, (discrete),

So $\text{E}(x) = \sum \text{x} \cdot P(\text{x} = \text{x}) = \$1.99$.
A: Before the drawing, someone offers you $x_0$ for your ticket; should you sell?

[ ] A: With $U(x)$ as your utility function, your expected gain if you keep the ticket is $E[U(x)]$; if for you $U(x) = x$ (utility = money) then

$E(U(x)) = 1.99$

Action 1 (sell): you gain $x_0$ for sure

Action 2 (keep): your expected utility is $E[U(x)]$

Under MEU you should sell if $U(x_0) > E[U(x)]$

If $U(x) = x$ for you then your optimal action is (sell if offered more than $1.99$).
Related but different problem.

On 13 Jan 2016, during the Powerball jackpot, the jackpot was $1.6 billion.

In your winnings, it is uncertain before the drawing.

Rejo calculation on p. 276: $E(E)$ is now $85.80 on a $2 ticket.

1. If $u(x) = x$ for you, under neutrality, is it rational to sell all your assets & buy as many lottery tickets as possible?

A: Yes, but that's a silly utility function; to be realistic, you'd have to subtract from $x$ the

new 1st row in table is

\[
\frac{1,600,000,000}{292,201,338} = 85.476
\]
A catalog of useful distributions

(Sch. 5) Case 2: Discrete

Bennoulli

$X \sim \text{Bennoulli}(p)$, $0 < p < 1$, if

$f_X(x) = p^x (1-p)^{1-x} I_{\{0,1\}}(x)

E(X) = p

Var(X) = p(1-p)

\text{Var}(X) = \frac{p(1-p)}{(1-p)}$

$\mu_X(t) = p e^t + (1-p)$ for all $-\infty < t < \infty$
If the $X_i$'s in $X_1, X_2, \ldots$ are i.i.d. Bernoulli $(p)$, then $(X_1, X_2, \ldots)$ are called Bernoulli trials with parameter $p$; if the sequence $(X_1, X_2, \ldots)$ is infinite this defines a Bernoulli (stochastic) process.

Binomial $X \sim \text{Binomial} \left( n, p \right)$ (i.e., $X$ follows the Binomial distribution with parameters $n$ (integer) and $0 < p < 1$)

\[
\begin{align*}
&f_X(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for} \quad x \in \{0, 1, \ldots, n\} \\
&\text{Support} (X)
\end{align*}
\]

Consequences $X_1, \ldots, X_n \sim \text{Bernoulli} (p)$

$\implies X = \sum_{i=1}^n X_i \sim \text{Binomial} (n, p)$
\( X \sim \text{Binomial} (n, p) \)  \( E(X) = np \)  \( V(X) = np(1-p) \)

\[ \Psi(t) = \left[ \frac{pe^t + (1-p)}{1 - (pe^t + (1-p))} \right]^n \quad \text{for all} \quad -\infty < t < \infty \]

**Supreme Court Case**

**Case Study**

Cartaneda v. Partida (1977)

Grand juries in the U.S. judicial system have 18 catchment areas: everybody over living in the judicial district for that grand jury (a few other minor restrictions)

Hidalgo County, Texas

\( 2\frac{1}{2} \) yr period at issue in Supreme Court Case: 220 people called to serve on grand juries, but only 100 of them were Mexican-American

eligible pool was 79.1% Mexican-American

extreme southern border of TX with Mexico

Q: Prima facie case of discrimination?
Before this 21st yr period, let \( \mathcal{I} \) be your prediction of the number of Mexican-Americans among the 220 people.

If no discrimination:

\[ I \sim \text{Binomial} \left( 220, 0.791 \right) \]

\[ T_i = \text{theory} \]

\[ n \cdot p \]

\[ E \left( I \mid T_i \right) = (220)(0.791) = 174.0 \]

\[ SD \left( I \mid T_i \right) = \sqrt{n \cdot p \cdot (1-p)} = 6 \]

If you were expecting 174 give or take 6, would you be surprised to see 100? [A: You'd be astonished]

Frequentist statistical answer:

\[ P \left( I \leq 100 \mid T_i \right) = 8.0 \times 10^{-28} \]

\( T_i \) looks ridiculous.

Bayesian statistical answer:

Need to compute \( P \left( T_i \mid I = 100 \right) \), not the other way around (later).
A finite population has \( A \) elements of type 1 and \( B \) elements of type 2; total population size \( (A+B) \).

You choose \( n \) elements at random without replacement from this population (i.e., you take a simple random sample (SRS) of size \( n \)).

Let \( X = \) (number of elements of type 1 in your sample)

Recall (as noted in take-home Test 2, Problem 2) \( X \) follows the hypergeometric distribution with parameters \((A, B, n)\). As we saw in that problem, the PDF of \( X \) is
\[ f_X(x | A, B, n) = \binom{A}{x} \binom{B}{n-x} \frac{I[\max\{0, n-B\} \leq x \leq \min\{n, A\}]}{(A+B)} \]

for \((A, B, n)\) non-negative integers with \(n \leq A + B\)

**Consequences**

1. \(E(X) = n \frac{A}{A + B}\)

2. \(V(X) = n \frac{A}{(A+B)(A+B+1)} \left( \frac{B}{A+B} \right) \left( \frac{A+B-n}{A+B-1} \right) \)

Note that if your sampling had been with replacement (i.e., you take an IID sample), \(X\) would have been binomial with the same value of \(n\) and \(p = \frac{A}{A+B}\) in that case. \(E(X) = n \frac{A}{A+B}\) and

\[ V(X) = np(1-p) = n \left( \frac{A}{A+B} \right) \left( \frac{B}{A+B} \right) \]
If you let $T = (A + B)$ be the total # of elements in the population,

<table>
<thead>
<tr>
<th>Sampling Method</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>with repl. (IID)</td>
<td>$n \left( \frac{A}{A+B} \right)$</td>
<td>$n \left( \frac{A}{A+B} \right) \left( \frac{B}{A+B} \right)$</td>
</tr>
<tr>
<td>without repl. (SRS)</td>
<td>$n \left( \frac{A}{A+B} \right)$</td>
<td>$n \left( \frac{A}{A+B} \right) \left( \frac{B}{A+B} \right) \left( \frac{T-n}{T-1} \right)$</td>
</tr>
</tbody>
</table>

$0 \leq \frac{T-n}{T-1} \leq 1$ is called the finite population correction.

3 special cases worth considering:

1. $(n=1)$ $\alpha = 1 \iff$ SRS = IID with only 1 element sampled.
2. $(n=T)$ $\alpha = 0 \iff$ If you exhaust the entire population with SRS, you have no uncertainty left.
many rvs that represent counts of occurrences of events in time intervals of fixed length have $VTHR > 1$.

The Poisson & Binomial distributions both count the number of "successes" in a process unfolding in time, so it should not be surprising to find out that these 2 dist. are related:

when ($n$ is large, $p$ is close to 0), $\text{Binomial}(n,p) \approx \text{Poisson}(np)$.

Theorem $n$ positive integer, $0 < p < 1$, $\xi \sim \text{Binomial}(n,p)$, $\lambda = np$, $\xi \sim \text{Poisson}(\lambda)$. Choose any sequence
A poisson process with rate $\lambda$ per unit time, is a stochastic process with the property:

(a) The number of arrivals in every interval of time is independent
(b) As time goes to zero, the probability $P(n, (0,t]) = \frac{e^{-\lambda t} (\lambda t)^n}{n!}$
(c) Time of first arrival is a non-overlapping (Markov chain) process with

$$f(x|\lambda, n) \to \text{as } n \to \infty$$
An organism called *Cryptosporidium* that's capable of getting into the public drinking water supply at one stage in their life cycle are called oocysts. They can make people sick at a concentration of only 1 oocyst per 5 liters = 1.3 gallons of water.

One problem is that it can be hard to detect these oocysts with water filtration. Suppose that, in the water supply of your city, oocysts occur according to a Poisson process with rate 2 oocysts per liter, and that the filtering system your water utility company uses can capture all the oocysts in a water sample but only has
probability \( p \) of detecting each oocyst that's actually there. \( (\& \text{counting events are independent}) \)

Set \( Y = \# \text{ oocysts in } t \text{ liters of water} \)

and \( \xi_i = \begin{cases} 1 & \text{if oocyst } i \text{ gets counted} \\ 0 & \text{else} \end{cases} \)

Then \( (\xi | Y = y) \sim \text{Binomial} (y, p) \)

The law of total probability:

\[
\begin{align*}
\mathbb{P}(X = x) &= \mathbb{P}(X = x) = \sum_{y=0}^{\infty} \mathbb{P}(Y = y) \mathbb{P}(X = x | Y = y) \\
&= \sum_{y=0}^{\infty} \binom{y}{x} p^x (1-p)^{y-x} \\
&= \frac{(2t)^x}{x!} e^{-2t} \quad \text{for } y > 0, x = 0, 1, \ldots
\end{align*}
\]

in which \( \mathbb{P}(Y = y) = \frac{(2t)^y}{y!} e^{-2t} \) for \( y = 0, 1, \ldots \)